Microbiology (2004), **150**, 4045–4052 DOI 10.1099/mic.0.27413-0

Distribution and characterization of staphylococcal interspersed repeat units (SIRUs) and potential use for strain differentiation

K. J. Hardy,[1,2] D. W. Ussery,[3] B. A. Oppenheim[1] and P. M. Hawkey[1,2]

[1]West Midlands Public Health Laboratory, Health Protection Agency, Heartlands Hospital, Bordesley Green East, Birmingham B9 5SS, UK

[2]Division of Infection and Immunity, University of Birmingham, Birmingham B15 2TT, UK

[3]Center for Biological Sequence Analysis, Biocentrum-DTU, Building 208, The Technical University of Denmark, DK-2800 Denmark

Correspondence
K. J. Hardy
Katie.Hardy@heartsol.wmids.nhs.uk

Received 17 June 2004
Revised 12 August 2004
Accepted 31 August 2004

Variable-number tandem repeats (VNTRs) have been shown to be a powerful tool in the determination of evolutionary relationships and population genetics of bacteria. The sequencing of a number of *Staphylococcus aureus* genomes has allowed the identification of novel VNTR sequences in *S. aureus*, which are similar to those used in the study of the evolution of *Mycobacterium tuberculosis* clades. Seven VNTRs, termed staphylococcal interspersed repeat units (SIRUs), distributed around the genome are described, occurring in both unique and multiple sites, and varying in length from 48 to 159 bp. Variations in copy numbers were observed in all loci, within both the sequenced genomes and the UK epidemic methicillin-resistant *S. aureus* (EMRSA) isolates. Clonally related UK EMRSA isolates were clustered using SIRUs, which provided a greater degree of discrimination than multi-locus sequence typing, indicating that VNTRs may be a more appropriate evolutionary marker for studying transmission events and the geographical spread of *S. aureus* clades.

INTRODUCTION

A number of different tandem repeat sequences have been described in both prokaryotes and eukaryotes. All have a unique length and DNA sequence of the repeat unit, different copy numbers per genome, and different organization within the genome. Loci with repeat units of 6–100 bp are referred to as minisatellites, whilst loci with smaller repeat units are termed microsatellites (Jeffreys *et al.*, 1992; Tautz & Schlotterer, 1994). The tandem repeat nature of minisatellites has been studied extensively in eukaryotic genomes. The first bacterial species in which they were identified was *Mycobacterium tuberculosis*, being described as mycobacterial interspersed repeat units (MIRUs) (Supply *et al.*, 1997, 2000). These loci have a characteristic structure consisting of a variable number of near-identical repeated DNA sequences arranged consecutively. A general term for these repeats is variable-number tandem repeats (VNTRs), and true VNTRs have now been reported in *Bacillus anthracis*, *Legionella pneumophila*, *Pseudomonas aeruginosa*, *Salmonella enterica* and *Escherichia coli* O157 (Keim *et al.*, 2000; Liu *et al.*, 2003; Noller *et al.*, 2003; Onteniente *et al.*, 2003; Pourcel *et al.*, 2003). With the exception of *M. tuberculosis*, these studies have been confined to comparisons of molecular strain differentiation techniques.

Staphylococcus aureus, and in particular methicillin-resistant *S. aureus* (MRSA), is an important nosocomial pathogen. In order to understand the molecular evolution of different strains in relation to important physiological properties such as pathogenicity and survival, evolutionary markers are needed. Multi-locus sequence typing (MLST) examines the slowly evolving genomic core, and is a useful tool for defining chromosomal ancestry (Enright *et al.*, 2000). Through use of this technique and others, the population structure of *S. aureus* has been studied, and it has been shown to consist of a limited number of clones in comparison to other bacterial species. Investigations into the evolution of MRSA have shown that horizontal transfer of the SCCmec complex, containing the *mecA* gene, has occurred on numerous occasions (Enright *et al.*, 2002). Despite this, the genetic backgrounds of hospital MRSA isolates are narrow in comparison to community methicillin-sensitive *S. aureus* (MSSA) and the recently described community MRSA (Enright *et al.*, 2002). Therefore, more rapidly evolving genetic markers are needed to study hospital MRSA.

Abbreviations: EMRSA, epidemic methicillin-resistant *Staphylococcus aureus*; MIRU, mycobacterial interspersed repeat unit; MLST, multi-locus sequence typing; MRSA, methicillin-resistant *S. aureus*; MSSA, methicillin-sensitive *S. aureus*; SIRU, staphylococcal interspersed repeat units; VNTR, variable-number tandem repeats.

The sequences of seven *S. aureus* genomes are now available; the sequences of N315, MW2, Mu50, MSSA 476 and MRSA 252 are published (Baba *et al.*, 2002; Kuroda *et al.*, 2001; Holden *et al.*, 2004), whilst the sequences of NCTC 8325 and COL are currently being annotated (http://www.genome.ou.edu/staph.html and http://www.tigr.org/tdb/mdb/mdbinprogress.html respectively). Five of the strains (COL, MRSA 252, Mu50, N315 and MW2) are MRSA, whilst MSSA 476 and NCTC 8325 are methicillin sensitive. N315 and Mu50 are closely related strains that are hospital-acquired and vancomycin-intermediate-resistance respectively, whilst MRSA 252 is an epidemic hospital strain, predominant within the UK. MW2 is a hypervirulent community strain isolated in the USA, but it has been associated with MSSA 476, a community isolate from the UK. The sequencing of these genomes has allowed us to identify novel multiple tandem repeats that are similar to MIRUs in *M. tuberculosis*, and the following report describes the nature and distribution of these repeats, and their potential as a novel tool for understanding the micro-evolution of hospital MRSA.

METHODS

Tandem repeat locus identification. Three of the sequenced *S. aureus* genomes, MW2, N315 and MRSA 252, were searched for the presence of tandem repeats using the tandem repeat finder program (http://c3.biomath.mssm.edu/trf.html).

The specified search criteria were that there must be >94 % conservation between the repeat units, and that the repeat unit must be between 45 and 200 bp in length. A total of 26 tandem repeats across the three genomes fitted these criteria, with seven identified in EMRSA 16, 13 in MW2 and six in N315. The presence of these repeat units, and the degree of similarity of the flanking regions in the other sequenced *S. aureus* genomes, were investigated using BLAST. Seven of the repeat units were present in all seven sequenced genomes, and they had a high degree of similarity in the flanking regions; these were designated staphylococcal interspersed repeat units (SIRUs) (Table 1). The presence of SIRU01, 05, 07, 13, 15, 16 and 21 was studied in a collection of the most widely distributed nosocomial epidemic MRSA (EMRSA) strains in the UK.

Primer design. Primers were designed to anneal within the conserved flanking regions outside of the repeat regions for all seven SIRUs. The length of flanking region included within the amplicon ranged from 103 to 260 bp. Primers were designed using the computer program Primer3 (http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi). In order to be able to amplify all DNA templates using the same reaction conditions, primer length and G+C content were kept as constant as possible. A common magnesium concentration and annealing temperature were identified using serial titrations.

Isolates and DNA preparation. A collection of the UK EMRSA types (1–16; NCTC), all of which have been previously described and typed using PFGE (Kerr *et al.*, 1990; Moore & Lindsay, 2002), were used to assess the discriminatory ability of the repeat units. DNA was extracted as described by Kumari *et al.* (1997).

PCR amplification and detection. Briefly, each PCR reaction (25 μl) contained 1× PCR Gold Buffer (Applied Biosytems), 200 μM each deoxynucleoside triphosphate (dNTP), 1 μM each flanking primer, 2 mM $MgCl_2$, 0·5 U AmpliTaq Gold (Applied Biosystems) and 2 μl DNA. Amplification was performed in a Multi-Block System (MBS) thermal cycler (ThermoHybaid). An initial denaturation temperature of 96 °C for 2 min was followed by 30 cycles of denaturation at 96 °C for 30 s, annealing at 55 °C, and elongation at 72 °C for 30 s. There was a final elongation step at 72 °C for 2 min. The amplicons were analysed on a 2 % (w/v) agarose gel (Invitrogen) for 2·5 h at 150 V, with 50 bp ladder size standards. Gels were stained using ethidium bromide, visualized with UV and photographed. The size of each amplicon was determined, and the number of repeats calculated, taking into account the size of the repeat unit and the flanking region.

DNA sequencing and PFGE. The DNA sequences of all seven SIRUs from EMRSA-16 were determined using Big Dye PCR reaction (Applied Biosystems), and analysed on a Primus HT 3700 DNA analyser by the functional Genomics Laboratory, University of Birmingham, UK. PFGE was carried out by using a standardized method with *Sma*I restriction endonuclease (Invitrogen) as described by Murchan *et al.* (2003).

Cluster analysis. SIRU and PFGE results were entered into BioNumerics (Applied Maths), analysed using Euclidian coefficient, and displayed via the unweighted pair group method using arithmetic averages (UPGMA).

RESULTS

We investigated seven sequenced *S. aureus* genomes for the presence of VNTRs. Seven VNTRs, termed SIRUs, were identified. Each SIRU was made up of repeated sequences, the number of which varied from strain to strain. The seven SIRUs were found to be distributed around the genome of *S. aureus* (Table 1).

Table 1. Genomic characteristics of SIRUs identified from sequenced *S. aureus* genomes

Locus no.	Genome tandem repeat finder performed on	Repeat length (bp)	Conservation (%)	G+C content (mol%)	Location on sequenced genome
01	EMRSA 16	55	98	40	384017–384123
05	EMRSA 16	60	98	36	2488212–2488383
07	EMRSA 16	56	95	46	2855668–2855783
13	MW2	64	100	30	1215469–1215625
15	MW2	131	99	32	1725259–1725542
16	MW2	159	98	35	1825755–1826239
21	N315	48	94	45	122940-123176

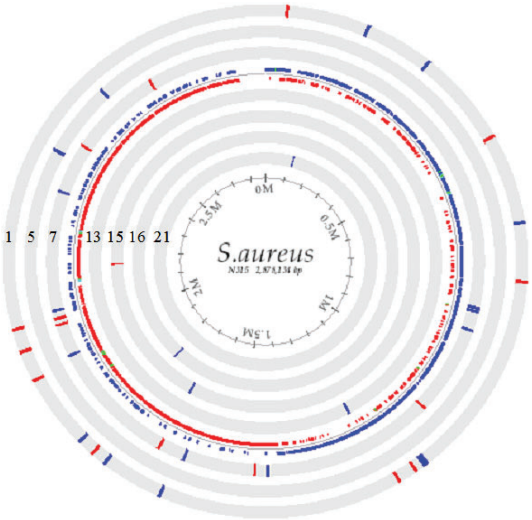

Fig. 1. Genome atlas of MRSA strain N315. The genomic position of each of the SIRUs is represented on the grey bands by red (positive DNA strand) and blue (negative DNA strands) markers. Between SIRU07 and 13, the genes on the forward and reverse strands are shown in red and blue respectively. In the centre circle, the number of bases around the genome is shown; 0M, the origin of replication.

Occurrence of repetitive sequences in sequenced genomes

The presence and position of the seven SIRUs were investigated in each of the seven sequenced *S. aureus* genomes. Interestingly two different phenomena were observed, in that four of the SIRUs were present only once on the genome (single-site SIRUs), whilst three SIRUs were present several times around the genome (multiple-site SIRUs) (Fig. 1). The positions of SIRUs 13, 15, 16 and 21, which occurred only once, were different relative to each other around the genome (Fig. 1). The positioning of single-site SIRUs remained constant between different genomes. These four single-site SIRUs had a very high percentage similarity in both the 3′ and 5′ flanking sequences. The multiple-site SIRUs 01 and 07 occurred at multiple sites around the genome. Variation in the number of the repeat units present at each site on the genome was observed, as well as some variation in the repeat sequence between sites. However, the flanking regions of each of the multiple-site SIRUs present on the same genome had very little similarity to each other. Similarity in flanking regions for specific sites was observed from genome to genome. The multiple-site SIRU that was selected for analysis contained the repeat units with the highest sequence homology, and all strains had flanking regions with high similarity. SIRU05 was present in two places on the genome. Both MW2 and MSSA 476 contained the repeat unit, and had a flanking region with 100 % similarity at the 3′ end, but the flanking region at the 5′ end was significantly different.

Comparison of numbers of repeat units and DNA sequences of SIRUs between different sequenced genomes

Length polymorphisms were seen at all loci (Table 2). The number of repeat units present in a genome was calculated on full repeat units only. SIRU01, 05, 15 and 16 all had full repeats (see Fig. 2 for SIRU05). When partial repeats were present, they were included in the flanking region for determination of the number of repeat units. For example, in SIRU15, an extra 22 bp of the repeat unit was present in all genomes, but in the calculation this was included in the flanking region (Fig. 3). A more complicated situation was observed in SIRU13 (Fig. 4), which had half-repeats present in all genomes. COL and NCTC 8325 had only half a repeat unit, and were therefore classed as having no full repeats in the calculations (Table 1).

There was a high level of sequence identity between the different sequenced genomes within the repeat units, with a few exceptions. The highest degrees of sequence identity in all repeat units were seen in SIRU05, 13 and 21. In contrast, COL, NCTC 8325 and MRSA 252 had a repeat unit

Table 2. Complete repeat units in the seven *S. aureus* sequenced genomes as shown by sequence alignment

Locus	Sequenced genome						
	NCTC 8325	COL	MW2	Mu50	N315	MRSA 252	MSSA 476*
1	3	4	1	2	2	2	1
5	6	6	5	3	3	3	5
7	3	2	3	1	1	2	–
13	0	0	2	3	3	1	1
15	1	1	2	1	1	2	2
16	4	4	3	3	3	2	1
21	5	5	4	5	5	5	4
Allelic profile	3630145	4620145	1532234	2313135	2313135	2321225	15–1214

*–, The flanking region and the repeat unit were not present.

```
SIRU05                                   [C TTTAACTTAA GTTATTAGAG CCTCTTATGC AGTTGCTCAG TCAACTGTAT ACCTTTTGC][C TTTAACTTAA GTTATTAGAG CCTCTTATGC
COL          AAGATTATTT ATATGGACTT TTAGAGGCT[C .......... .......... .......... .......... .......... .........][. .......... .......... ..........
NCTC8325     .......... .......... .........[. .......... .......... .......... .......... .......... .........][. .......... .......... ..........
MRSA252      .......... .......... .........[. .......... .......... .......... .......... .......... .........][. .......... .......... ..........
Mu50         .......... .......... .........[. .......... .......... .......... ......GT.. .......... .......T.][. .......... .......... ..........
N315         .......... .......... .........[. .......... .......... .......... ......GT.. .......... .......T.][. .......... .......... ..........
MSSA476      .......... .......... .........[. .......... .......... .......... ......GT.. .......... ......CA.][. .......... .......... ..........
MW2          .......... .......... .........[. .......... .......... .......... ......GT.. .......... ......C..][. .......... .......... ..........

SIRU05       AGTTGCTCAG TCAACTGTAT ACCTTTTGC][C TTTAACTTAA GTTATTAGAG CCTCTTATGC AGTTGCTCAG TCAACTGTAT ACCTTTTGC][C TTTAACTTAA GTTATTAGAG CCTCTTATGC
COL          .......... .......... .........][. .......... .......... .......... ......GT.. .......... .........][. .......... .......... ..........
NCTC8325     .......... .......... .........][. .......... .......... .......... .......... .......... .........][. .......... .......... ..........
MRSA252      .......... .......... .........][. .......... .......... .......... .......... .......... .........][- ---------- ---------- ----------
Mu50         .......... .......... .......T.][. .......... .......... .......... .......... ATC.TC.... .........][- ---------- ---------- ----------
N315         .......... .......... .......T.][. .......... .......... .......... .......... ATC.TC.... .........][- ---------- ---------- ----------
MSSA476      .......... .......... .........][. .......... .......... .......... .......... .......... .........][. .......... .........A ..........
MW2          .......... .......... .........][. .......... .......... .......... .......... .......... .........][. .......... .........A ..........

SIRU05       AGTTGCTCAG TCAACTGTAT ACCTTTTGC][C TTTAACTTAA GTTATTAGAG CCTCTTATGC AGTTGCTCAG TCAACTGTAT ACCTTTTGC][C TTTAACTTAA GTTATTAGAG CCTCTTATGC
COL          .......... .......... .........][. .......... .........T. .......... ......GT.. .......... .........][. .......... .......... ..........
NCTC8325     .......... .......... .........][. .......... .........T. .........T ......GT.. .......... .........][. .......... .......... ..........
MRSA252      ---------- ---------- ---------][- ---------- ---------- ---------- ---------- ---------- ---------][- ---------- ---------- ----------
Mu50         ---------- ---------- ---------][- ---------- ---------- ---------- ---------- ---------- ---------][- ---------- ---------- ----------
N315         ---------- ---------- ---------][- ---------- ---------- ---------- ---------- ---------- ---------][- ---------- ---------- ----------
MSSA476      .......... .......... .........][. .......... .......... .......... .......... .......... ...CAA.T.][- ---------- ---------- ----------
MW2          .......... .......... .........][. .......... .......... .......... .......... .......... ...CAA.T.][- ---------- ---------- ----------

SIRU05       AGTTGCTCAG TCAACTGTAT ACCTTTTGC]
COL          ......G... ATC.TC.... .AAAA..AAT] GACGTCATTT
NCTC8325     ......G... ATC.TC.... .........] ..........
MRSA252      ---------- ---------- ---------] ..........
Mu50         ---------- ---------- ---------] ..........
N315         ---------- ---------- ---------] ..........
MSSA476      ---------- ---------- ---------] C..AA.CAA.
MW2          ---------- ---------- ---------] C..AA.CAA.
```

Fig. 2. Alignment of SIRU05 60 bp repeat units in the seven sequenced genomes. Deletion of a repeat is represented by a dash; a nucleotide present in all repeats is represented by a dot; repeat units are located within square brackets; flanking regions are underlined.

that was identical in length in SIRU15, but differed by 26 % from the consensus sequence (Fig. 3). N315 and Mu50 have been shown to be clonal (Kuroda *et al.*, 2001), and investigations of their SIRUs show that they have the same number of repeat units in all loci. They also contain the same base changes in all repeat units. Similarity in the base changes observed in the repeat units has been seen in other SIRU repeat units, for example in SIRU16 of NCTC 8325 and COL.

The repeat units were true tandem repeats in all SIRUs, with no intervening sequence between the repeats, apart from SIRU01. SIRU01 had either 1 or 2 bp between the first and second repeats in all genomes. When sizing a fragment using a gel-based method, the presence of 1 or 2 bp does not change the apparent number of copies present.

Position of repeat

The published *S. aureus* sequences have been annotated, and we were able to establish the position of the SIRUs

```
SIRU15                                  [ATTA GGGGCTCTTA TGCAGTTGCT TTTAAATTTA CTTTACTCAA ATTGCTGATA TTTCCGAATA CCAAATATGA TACTTTTTTT GCTTTTTATA
COL          TTAGAGCCCC TTATATATTG TAAATA[.... .......... .......A.. ....T.AA.C T..GCT.T.. .A.AC.T... ....TC.GC. ....T.G..G ...AG...A. T....AA.CG
NCTC8325     .......... .......... ......[.... .......... .......A.. ....T.AA.C T..GCT.T.. .A.AC.T... ....TC.GC. ....T.G..G ...AG...A. T....AA.CG
MRSA252      .......... .......... ......[.... .......... .......... ....T.AA.C T..GCT.T.. .A.AC.T... ....TC.AC. ....T.G..G ...AG...A. T....AA.CG
Mu50         .......... .......... ......[.... .......... .......... .......... .......... .......... .......... .......... .......... ..........
N315         .......... .......... ......[.... .......... .......... .......... .......... .......... .......... .......... .......... ..........
MW2          .......... .......... ......[.... .......... .......... .......... .......... .......... .......G.. .......... .......... ..........
MSSA476      .......... .......... ......[.... .......... .......... .......... .......... .......... .......G.. .......... .......... ..........

SIRU15       CTGCAACTGT ATTCTTTTGT CCCC-TTATA GTTTAGCG][ATTAGGGG CTCTTATGCA GTTGCTTTTA AATTTACTTT ACTCAAATTG CTGATATTTC CGAATACCAA ATATGATACT
COL          .......... ...C....TA ....C..... ......TG]{-------- ---------- ---------- -------------------- ---------- ---------- ----------
NCTC8325     .......... ...C....TA ....C..... ......TG]{-------- ---------- ---------- -------------------- ---------- ---------- ----------
MRSA252      .......... ...C....TA G...C..... TA..GTAA][........ .......... .......... .G...................... .......... .......... ..TC......
Mu50         .......... .......... ....-..... ........]{-------- ---------- ---------- -------------------- ---------- ---------- ----------
N315         .......... .......... ....-..... ........]{-------- ---------- ---------- -------------------- ---------- ---------- ----------
MW2          .......... .......... ....-..... ........][........ .......... .......... .......... .......... .......... .......... ..........
MSSA476      .......... .......... ....-..... ........][........ .......... .......... .......... .......... .......... .......... ..........

SIRU15       TTTTTTGCTT TTTATACTGC AACTGTATTC TTTTGTCCCC TTATAGTT-A GCG][ATTAGGG GCTCTTATGC AGTTG]
COL          ---------- ---------- ---------- ---------- ---------- ---][....... ........A. .....]-ACAT TTAGTATAA CGTCTAAAAT
NCTC8325     ---------- ---------- ---------- ---------- ---------- ---][....... ........A. .....]C.... .......... ..........
MRSA252      ....--.... .......... .......... C......... ........T. ...][....... .......... .....]C.... .......... T.....T...
Mu50         ---------- ---------- ---------- ---------- ---------- ---][....... .......... .....]T.... .......... ..........
N315         ---------- ---------- ---------- ---------- ---------- ---][....... .......... .....]T.... .......... ..........
MW2          .......... .......... .......... .......... ........T.....][....... .......... .....]T.... .......... ..........
MSSA476      .......... .......... .......... .......... ........T. ...][....... .......... .....]T.... .......... ..........
```

Fig. 3. Alignment of SIRU15 131 bp repeat units in the seven sequenced genomes. Deletion of a repeat is represented by a dash; a nucleotide present in all repeats is represented by a dot; repeat units are located within square brackets; flanking regions are underlined.

```
SIRU13                                  [TAAGAAA CACTAATCAA TAAATTGATA AGAAGATTAG AGGAACGCTT GAGCTAAAGC TCATGCA]{TAA GAAACACTAA TCAATAAATT GATAAGAAGA
COL          GATAAGGTTC CAGAATTTCT CCC[....... .......... .......... ..TG------ ---------- ---------- -------]{--- ---------- ---------- ----------
NCTC8325     .......... .......... ...[....... .......... .......... ..TG------ ---------- ---------- -------]{--- ---------- ---------- ----------
MRSA252      .......... .......... ...[....... ..G....... .......... .CT.ATAAGC G.A-....C. .......... .......]{... ......G... .......... .....CTG--
Mu50         .......... .......... ...[....... .......... .......... .......... .......... .......... ...A.AT]{A.. ..CG..GA.. .......... .....CTG.G
N315         .......... .......... ...[....... .......... .......... .......... .......... .......... ...A.AT]{A.. ..CG..GA.. .......... .....CTG.G
MW2          ......... G..-----.. ..-[...-... .......... .......... .......... .......... .......... .......]{... .......... .......... ..........
MSSA476      .......... G..---..-. ...[C...... .......... .......... .......... .......... .......... .......]{... .......... .......... ......TG--

SIRU13       TTAGAGGAAC GCTTGAGCTA AAGCTCATGC A]{TAAGAAACA CTAATCAATA AATTGATAAG AAGATTAGAG GAACGCTTGA GCTAAAGCTC ATGCA]{TAAGA AACACTAATC AATAAATTGA
COL          ---------- ---------- ---------- -]{--------- ---------- ---------- ---------- ---------- ---------- -----]{----- ---------- ----------
NCTC8325     ---------- ---------- ---------- -]{--------- ---------- ---------- ---------- ---------- ---------- -----]{----- ---------- ----------
MRSA252      ---------- ---------- ---------- -]{--------- ---------- ---------- ---------- ---------- ---------- -----]{----- ---------- ----------
Mu50         A......T.. ..C....... .......A.A T]{A....CG.. G......... .........C TG.GA..... .T..A.C... .......... .....]{..... .......... ..........
N315         A......T.. ..C....... .......A.A T]{A....CG.. G......... .........C TG.GA..... .T..A.C... .......... .....]{..... .......... ..........
MW2          .......... .......... .......... .]{......... .......... .......... TG-------- ---------- ---------- -----]{----- ---------- ----------
MSSA476      ---------- ---------- ---------- -]{--------- ---------- ---------- ---------- ---------- ---------- -----]{----- ---------- ----------

SIRU13       TAAGAA
COL          ------TTTC TAAAATTTCT ACTTGTTTTT TAGAATTTAA
NCTC8325     ------.... .......... .......... ..........
MRSA252      ------.... .......... .......... ..........
Mu50         ....TG.... .......... .......... ..........
N315         ....TG.... .......... .......... ..........
MW2          ------.... .......... .......... ..........
MSSA476      ------.... .......... .......... ..........
```

Fig. 4. Alignment of SIRU13 64 bp repeat units in the seven sequenced genomes. Deletion of a repeat is represented by a dash; a nucleotide present in all repeats is represented by a dot; repeat units are located within square brackets; flanking regions are underlined.

in relation to coding and non-coding regions. Five of the seven SIRUs were in non-coding regions. SIRU13 was in the coding region for MW1681, a hypothetical protein. SIRU21 was in the coding region for protein A, *spa*. This region has been previously used for epidemiological typing (Koreen *et al.*, 2004). Variation in the number of *spa* repeats is based on a repeat size of 24 bp; our software identified a repeat of 48 bp. SIRU15 and 16 were in non-coding regions, but were close to the *obg* and *hmrB* genes respectively.

Use of repeats in defining the genetic relatedness of EMRSA types

PCR was performed on the EMRSA isolates from the UK which have been previously shown to differ by varying degrees by PFGE and MLST, to evaluate the potential discriminatory ability of the loci. Of the 16 isolates typed, 12 different SIRU profiles were observed, with different degrees of variation within the SIRU profile. Length polymorphisms of the amplified products were seen in all loci, with the greatest variations being in SIRU05, 13 and 21 (Table 3). On amplification, a single band was obtained for all reactions, confirming the specificity of the primers, and reproducibility was demonstrated, with a product of the same length being obtained on repetitive amplification of the same isolates. The sequence data obtained from sequencing all seven SIRUs from EMRSA16 and comparing them to the published EMRSA16 genome (MRSA 252) showed 100 % concordance. Further confirmation of the correct amplification of each of the SIRUs was demonstrated by the fact that all our amplicon sizes fell within the predicted size for an exact number of repeats.

SIRU typing had the ability to discriminate between the EMRSA strains (Fig. 5). A cluster of ten EMRSA isolates containing five different SIRU profiles differed by a maximum of two loci, and were clustered into three sequence types (STs) by MLST. The STs contained within this cluster only differed at one locus, either *pta* (phosphate acetyltransferase) or *arcC* (carbamate kinase). For this set of ten isolates, the MLST types and SIRU profiles were not always in concordance. The remaining six EMRSA isolates had distinct SIRU profiles and MLST types, and a greater degree of variation within the loci was observed.

DISCUSSION

The value of VNTRs as a powerful tool in the investigation of both population genetics and evolutionary events has

Table 3. Number of complete repeat units identified in each of the UK hospital EMRSA isolates tested in the field trial

Isolate	Locus no.*							Allelic profile*
	01	13	15	16	21	05	07	
EMRSA 1	2	1	1	3	3	2	2	2113322
EMRSA 2	2	1	1	3	3	1	3	2113313
EMRSA 3	1	3	1	3	5	2	1	1313521
EMRSA 4	3	1	1	3	3	2	2	3113322
EMRSA 5	3	1	1	x	5	3	2	311x532
EMRSA 6	3	1	1	3	3	1	3	3113313
EMRSA 7	3	1	1	3	3	2	2	3113322
EMRSA 8	3	1	1	4	5	7	2	3114572
EMRSA 9	3	1	1	3	3	2	2	3113322
EMRSA 10	3	1	1	3	7	8	2	3113782
EMRSA 11	2	1	1	3	3	2	2	2113322
EMRSA 12	3	1	1	3	4	1	3	3113413
EMRSA 13	3	1	1	3	3	1	3	3113313
EMRSA 14	3	1	1	3	3	1	3	3113313
EMRSA 15	1	5	0	3	8	x	2	15038x2
EMRSA 16	1	2	2	2	5	3	2	1222532

*x, locus in which no amplification was obtained.

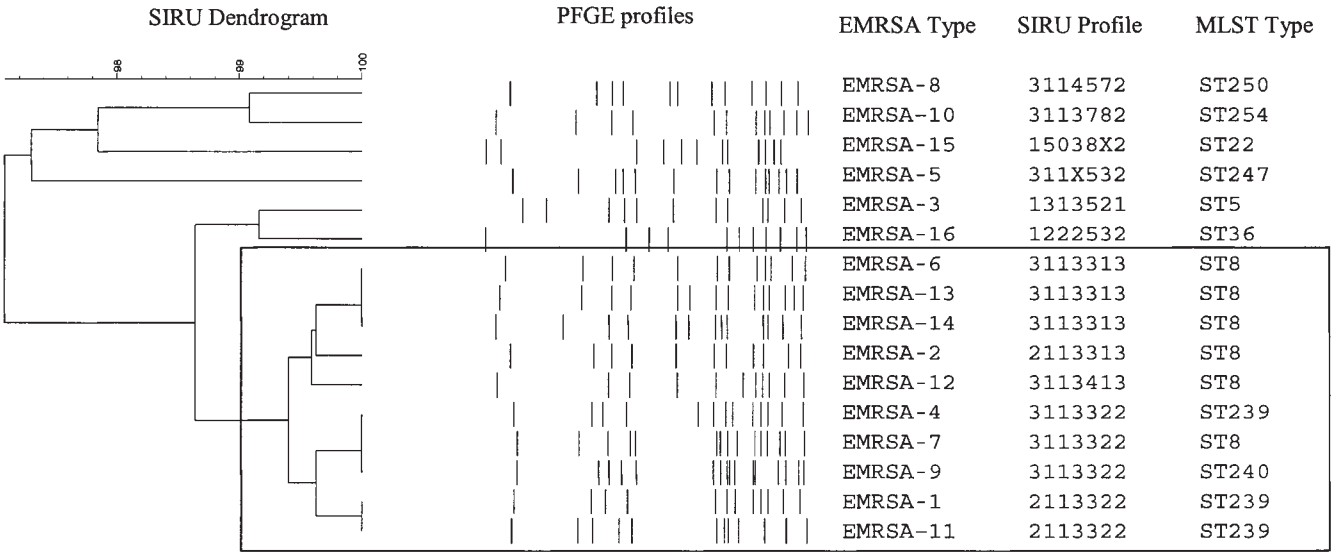

Fig. 5. Dendrogram deduced from the cluster analysis of the EMRSA strains. PFGE profiles, SIRU profiles and corresponding MLST types are shown. The box highlights SIRU profiles that differ from each other by a maximum of two loci.

been described by other authors, first in eukaryotes, and subsequently in prokaryotes (Jeffreys *et al.*, 1985; Frothingham *et al.*, 1998). The availability of multiple sequenced *S. aureus* genomes has allowed us to make a comparative study of the VNTRs, and their potential role in determining the evolution of UK hospital MRSA isolates.

Several of the eukaryotic VNTRs have been associated with important functional effects; for example, they influence the transcription of insulin, and are associated with increased risk of insulin-dependent diabetes and several cancers (Bennett *et al.*, 1995; Krontiris *et al.*, 1993). Within bacteria, they have been described within both coding and non-coding regions, and, in *Haemophilus influenzae*, the loss or gain of repeats has been associated with the phase variation that plays an important role in its pathogenicity (Van Belkum *et al.*, 1997). Although all of the SIRUs are in non-coding regions on the positive strand, the coding region for the protein A gene *spa* is on the complementary strand of SIRU21. Protein A is a cell-wall-anchored protein, and has been shown to have a possible role in the virulence of *S. aureus* infections; the effect of variable numbers of tandem repeats on the *spa* gene is not known.

In this study, the repeat units of the SIRUs differed in length, varying from 48 to 159 bp. All the repeat units at the different loci contained point mutations, and small insertions and deletions; this phenomenon was first described in the minisatellite in the human insulin gene, and has since been shown to be the case in all VNTR loci described, both eukaryotic and prokaryotic (Mazars *et al.*, 2001; Owerbach & Aagaard, 1984). Like the VNTRs described in other bacterial genomes, and unlike those in eukaryotes, the repeat units in *S. aureus* are not G+C rich (Le Fleche *et al.*, 2001).

Variations in copy number were observed in all loci, within both the sequenced *S. aureus* genomes and the EMRSA strains used in the field trial. Variation in repeat number within some of the loci was greater (SIRU05, 13 and 21), and therefore demonstrated a greater discriminatory power; this is similar to the phenomenon described in *M. tuberculosis* (Supply *et al.*, 2000). At one time, the variation in copy number was thought to be due to an unequal crossing-over exchange between homologous chromosomes, but Wolff and colleagues demonstrated that there was no exchange of flanking DNA when a repeat unit was lost (Wolff *et al.*, 1988). Other theories have involved misalignment mutational mechanisms, or changes in allele number arising from mutation initiated by double-strand breaks (Lovett, 2004; Paques *et al.*, 1998). Also, as described in *M. tuberculosis*, within some loci, certain allele copy numbers are more frequent, but the presence of all the loci adds to the discriminatory power of the technique (Mazars *et al.*, 2001). The factors responsible for the difference in mutation rates are not clear, but mutation rates may be promoted by factors such as short repeat units, high copy number arrays and repeat sequence homogeneity (Rocha, 2003). Interestingly the loci with the longest repeat units in *S. aureus* have the least variation in copy number.

Partial repeats of varying length are present at the end of all the loci; this has been described in *M. tuberculosis*, where in H37Rv, for example, the ETR-B locus contains three complete copies of the 57 bp repeat, and also an additional eight bases corresponding to the beginning of another tandem repeat (Frothingham & Meeker-O'Connell, 1998). The significance of these partial repeats is not known.

The flanking regions of each individual locus had a high

sequence homology with other sequenced genomes, but they did have point mutations. The availability of multiple sequenced genomes meant that our study did not suffer from the limitations of some other studies, in that we were able to design primers in a homologous region (Pourcel *et al.*, 2003).

The epidemiology and evolution of MRSA have been studied using several different typing techniques. The most widely used epidemiological tool is PFGE, which has a high discriminatory power. However, in some instances, the genetic markers that PFGE examines undergo rapid evolutionary change, and strains that are clonally related may produce diverse banding patterns (Blanc *et al.*, 2002). MLST examines the more slowly evolving genomic core, and is useful in the determination of evolutionary events. It has been applied to defining the evolutionary relationships of MRSA on a worldwide basis, but the applicability to local epidemiological settings has not been determined. The EMRSA isolates examined in our study were clustered into the same groups using PFGE, MLST and SIRU. PFGE showed the greatest discrimination, but SIRU typing had greater discrimination than MLST, being able to subdivide the isolates with the same MLST profiles. With its ability to cluster clonally related isolates together, but provide a greater degree of discrimination than MLST, SIRU indicates that the speed of change of VNTRs may be a more appropriate evolutionary clock for studying transmission events, as demonstrated in *M. tuberculosis* (Hawkey *et al.*, 2003). Recently 15 VNTR sequences were used to examine the phylogeny of a diverse collection of geographically and temporally separated strains of *M. tuberculosis* (Sola *et al.*, 2003). When compared with other molecular methods, variation in VNTR loci not only supported the accepted phylogenic structure, but also corroborated a hypothetical evolutionary scenario for the early divergence of *M. tuberculosis* from other closely related species. Sabat and colleagues reported the presence of several VNTR loci in *S. aureus* (not MRSA), described as multiple-locus VNTR analysis (Sabat *et al.*, 2003). The numbers of copies present were not determined, and the primer pairs were multiplexed to give a fingerprint pattern. The gene *spa*, encoding Protein A, contains a VNTR sequence within it, which we have identified as SIRU21. Variation in copy number and DNA sequence within the repeat units of *spa* has been used as a genetic marker in several studies, but it restricts the evolutionary studies to examining only one region of the genome (Koreen *et al.*, 2004). With our description of SIRUs, we expect that use can be made of the SIRU loci for enabling phylogenetic reconstruction of *S. aureus*, similar to the case with VNTR sequences and *M. tuberculosis*.

ACKNOWLEDGEMENTS

The authors wish to thank Wyeth for the support of K. J. Hardy, the Danish Center for Scientific Computing, whose grant supported part of this work, and BBSRC grant no. 6/JIS 13209 for sequencing.

REFERENCES

Baba, T., Takeuchi, F., Kuroda, M. & 11 other authors (2002). Genome and virulence determinants of high virulence community-acquired MRSA. *Lancet* **359**, 1819–1827.

Bennett, S. T., Lucassen, A. M., Gough, S. C. & 7 other authors (1995). Susceptibility to human type 1 diabetes at IDDM2 is determined by tandem repeat variation at the insulin gene minisatellite locus. *Nat Genet* **9**, 284–292.

Blanc, D. S., Francioli, P. & Hauser, P. M. (2002). Poor value of pulsed-field gel electrophoresis to investigate long-term scale epidemiology of methicillin-resistant *Staphylococcus aureus*. *Infect Genet Evol* **2**, 145–148.

Enright, M. C., Day, N. P., Davies, C. E., Peacock, S. J. & Spratt, B. G. (2000). Multilocus sequence typing for characterization of methicillin-resistant and methicillin-susceptible clones of *Staphylococcus aureus*. *J Clin Microbiol* **38**, 1008–1015.

Enright, M. C., Robinson, D. A., Randle, G., Feil, E. J., Grundmann, H. & Spratt, B. G. (2002). The evolutionary history of methicillin-resistant *Staphylococcus aureus* (MRSA). *Proc Natl Acad Sci U S A* **99**, 7687–7692.

Frothingham, R. & Meeker-O'Connell, W. A. (1998). Genetic diversity in the *Mycobacterium tuberculosis* complex based on variable numbers of tandem DNA repeats. *Microbiology* **144**, 1189–1196.

Hawkey, P. M., Smith, E. G., Evans, J. T., Monk, G., Bryan, H., Mohamed, H. & Bardhan M.& Pugh, R. N. (2003). Mycobacterial interspersed repetitive unit typing of *Mycobacterium tuberculosis* compared to IS6110-based restriction fragement length polymorphism analysis for investigation of apparently clustered cases of tuberculosis. *J. Clin. Microbiol* **41**, 3514–3520.

Holden, M. T., Feil, E. J., Lindsay, J. A. & 42 other authors (2004). Complete genomes of two clinical *Staphylococcus aureus* strains: evidence for the rapid evolution of virulence and drug resistance. *Proc Natl Acad Sci U S A* **101**, 9786–9791.

Jeffreys, A. J., Wilson, V. & Thein, S. L. (1985). Individual-specific 'finger-prints' of human DNA. *Nature* **316**, 76–79.

Jeffreys, A. J., Wilson, V. & Thein, S. L. (1992). Hypervariable 'minisatellite' regions in human DNA. *Biotechnology* **24**, 467–472.

Keim, P., Price, L. B., Klevytska, A. M., Smith, K. L., Schupp, J. M., Okinaka, R., Jackson, P. J. & Hugh-Jones, M. E. (2000). Multiple-locus variable-number tandem repeat analysis reveals genetic relationships within *Bacillus anthracis*. *J Bacteriol* **182**, 2928–2936.

Kerr, S., Kerr, G. E., Makintosh, C. A. & Marples, R. R. (1990). A survey of methicillin-resistant *Staphylococcus aureus* affecting patients in England and Wales. *J Hosp Infect* **16**, 35–48.

Koreen, L., Ramaswamy, S. V., Graviss, E. A., Naidich, S., Musser, J. M. & Kreiswirth, B. N. (2004). *spa* typing method for discriminating among *Staphylococcus aureus* isolates: implications for use of a single marker to detect genetic micro- and macrovariation. *J Clin Microbiol* **42**, 792–799.

Krontiris, T. G., Devlin, B., Karp, D. D., Robert, N. J. & Risch, N. (1993). An association between the risk of cancer and mutations in the HRAS1 minisatellite locus. *N Engl J Med* **329**, 517–523.

Kumari, D. N., Keer, V., Hawkey, P. M., Parnell, P., Joseph, N., Richardson, J. F. & Cookson, B. (1997). Comparison and application of ribosome spacer DNA amplicon polymorphisms and pulsed-field gel electrophoresis for differentiation of methicillin-resistant *Staphylococcus aureus* strains. *J Clin Microbiol* **35**, 881–885.

Kuroda, M., Ohta, T., Uchiyama, I. & 34 other authors (2001). Whole genome sequencing of methicillin-resistant *Staphylococcus aureus*. *Lancet* **357**, 1225–1240.

Le Fleche, P., Hauck, Y., Onteniente, L. & 7 other authors (2001). A tandem repeats database for bacterial genomes: application to the genotyping of *Yersinia pestis* and *Bacillus anthracis*. *BMC Microbiol* **1**, 2.

Liu, Y., Lee, M. A., Ooi, E. E., Mavis, Y., Tan, A. L. & Quek, H. H. (2003). Molecular typing of *Salmonella enterica* serovar typhi isolates from various countries in Asia by a multiplex PCR assay on variable-number tandem repeats. *J Clin Microbiol* **41**, 4388–4394.

Lovett, S. T. (2004). Encoded errors: mutations and rearrangements mediated by misalignment at repetitive DNA sequences. *Mol Microbiol* **52**, 1243–1253.

Mazars, E., Lesjean, S., Banuls, A. L., Gilbert, M., Vincent, V., Gicquel, B., Tibayrenc, M., Locht, C. & Supply, P. (2001). High-resolution minisatellite-based typing as a portable approach to global analysis of *Mycobacterium tuberculosis* molecular epidemiology. *Proc Natl Acad Sci U S A* **98**, 1901–1906.

Moore, P. C. L. & Lindsay, J. A. (2002). Molecular characterisation of the dominant UK methicillin-resistant *Staphylococcus aureus* strains, EMRSA-15 and EMRSA-16. *J Med Microbiol* **51**, 516–521.

Murchan, S., Kaufmann, M. E., Deplano, A. & 21 other authors (2003). Harmonization of pulsed-field gel electrophoresis protocols for epidemiological typing of strains of methicillin-resistant *Staphylococcus aureus*: a single approach developed by consensus in 10 European laboratories and its application for tracing the spread of related strains. *J Clin Microbiol* **41**, 1574–1585.

Noller, A. C., McEllistrem, M. C., Pacheco, A. G., Boxrud, D. J. & Harrison, L. H. (2003). Multilocus variable-number tandem repeat analysis distinguishes outbreak and sporadic *Escherichia coli* O157 : H7 isolates. *J Clin Microbiol* **41**, 5389–5397.

Onteniente, L., Brisse, S., Tassios, P. T. & Vergnaud, G. (2003). Evaluation of the polymorphisms associated with tandem repeats for *Pseudomonas aeruginosa* strain typing. *J Clin Microbiol* **41**, 4991–4997.

Owerbach, D. & Aagaard, L. (1984). Analysis of a 1963-bp polymorphic region flanking the human insulin gene. *Gene* **32**, 475–479.

Paques, F., Leung, W. Y. & Haber, J. E. (1998). Expansions and contractions in a tandem repeat induced by double-strand break repair. *Mol Cell Biol* **18**, 2045–2054.

Pourcel, C., Vidgop, Y., Ramisse, F., Vergnaud, G. & Tram, C. (2003). Characterization of a tandem repeat polymorphism *in Legionella pneumophila* and its use for genotyping. *J Clin Microbiol* **41**, 1819–1826.

Rocha, E. P. (2003). DNA repeats lead to the accelerated loss of gene order in bacteria. *Trends Genet* **19**, 600–603.

Sabat, A., Krzyszton-Russjan, J., Strzalka, W., Filipek, R., Kosowska, K., Hryniewicz, W., Travis, J. & Potempa, J. (2003). 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* **41**, 1801–1804.

Sola, C., Filliol, I., Legrand, E., Lesjean, S., Locht, C., Supply, P. & Rastogi, N. (2003). Genotyping of the *Mycobacterium tuberculosis* complex using MIRUs: association with VNTR and spoligotyping for molecular epidemiology and evolutionary genetics. *Infect Genet Evol* **3**, 125–133.

Supply, P., Magdalena, J., Himpens, S. & Locht, C. (1997). Identification of novel intergenic repetitive units in a mycobacterial two-component system operon. *Mol Microbiol* **26**, 991–1003.

Supply, P., Mazars, E., Lesjean, S., Vincent, V., Gicquel, B. & Locht, C. (2000). Variable human minisatellite-like regions in *the Mycobacterium tuberculosis* genome. *Mol Microbiol* **36**, 762–771.

Tautz, D. & Schlotterer, C. (1994). Simple sequences. *Curr Opin Genet Dev* **4**, 832–837.

Van Belkum, A., Scherer, S., van Leeuwen, W., Willemse, D., van Alphen, L. & Verbrugh, H. (1997). Variable number of tandem repeats in clinical strains of *Haemophilus influenzae*. *Infect Immun* **65**, 5017–5027.

Wolff, R. K., Nakamura, Y. & White, R. (1988). Molecular characterization of a spontaneously generated new allele at a VNTR locus: no exchange of flanking DNA sequence. *Genomics* **3**, 347–351.