The phylogeny of *Staphylococcus aureus* – which genes make the best intra-species markers?

Jessica E. Cooper and Edward J. Feil

Correspondence
Edward J. Feil
e.feil@bath.ac.uk

Department of Biology and Biochemistry, University of Bath, Claverton Down,
Bath BA2 7AY, UK

Received 21 October 2005
Revised 21 January 2006
Accepted 30 January 2006

The ability to make informed decisions on the suitability of alternative marker loci is central for population and epidemiological investigations. This issue was addressed using *Staphylococcus aureus* as a model population by generating nucleotide sequence data from 33 gene fragments in a representative sample of 30 strains. Supplementing the data with pre-existing multilocus sequence typing data, an intra-species tree based on ~17·8 kb of sequence was reconstructed and the goodness of fit of each individual gene tree was computed. No strong association was noted between gene function per se and phylogenetic reliability, but it is suggested that candidate loci should possess at least the average degree of nucleotide diversity for all genes in the genome. In the case of *S. aureus* this threshold is >1 % mean pairwise diversity.

INTRODUCTION

The influx of genomic and multilocus sequence data has transformed our understanding of bacterial evolution, and is set to revolutionize bacterial systematics and our view of what constitutes a bacterial ‘species’ (Gevers *et al.*, 2005). In particular, recent years have seen the rise of multilocus sequence typing (MLST) for epidemiological or population studies on single named species. These studies commonly involve the characterization of hundreds of isolates at a small number of gene loci, assumed to be a representative sample of the ‘core’ genome (‘housekeeping’ genes). Homologous recombination, the replacement of a gene with an orthologue from an unrelated lineage, may confound attempts at intra-species phylogenetic reconstruction or accurate typing (Feil *et al.*, 1999, 2000; Jolley *et al.*, 2000). The use of multiple (typically seven) loci in MLST is necessary to ‘buffer’ against this effect in single genes (Hanage *et al.*, 2005), and the employment of housekeeping genes is presumed to provide added insurance as there is no a priori reason to expect recombination to confer a selective advantage at such genes (Maiden *et al.*, 1998; Spratt & Maiden, 1999).

Whilst practical considerations dictate that candidate markers should be ubiquitous throughout the population

under consideration and present in single copy, other desirable criteria are not so clear-cut. For example, it is typically not possible to gauge which genes most closely reflect the underlying organismal phylogeny, or even if such a phylogeny exists (Baptiste *et al.*, 2005). Although genes encoding essential housekeeping functions are commonly viewed as the most reliable markers, the precise importance of gene function in predicting the utility of intra-species markers has not been systematically studied. Similarly, the optimal window of variation remains poorly defined, although it is clear that too little variation will result in poor resolution whereas too much will separate isolates that are very closely related.

It might be expected that genes encoding proteins which interact with the host or the external environment will be highly variable owing to strong diversifying selection, and as such be poor reflections of the underlying phylogeny. Two recent reports have compared the phylogenetic signal of MLST (housekeeping) genes in *Staphylococcus aureus* with those of highly variable genes encoding proteins putatively associated with the cell wall (Robinson *et al.*, 2005) or adhesins implicated to play a central role in host colonization and/or virulence (Kuhn *et al.*, 2006). Contrary to expectations, both investigations noted that highly variable genes were at least as informative for phylogenetic reconstruction as the slowly evolving housekeeping genes. These observations suggest that strong diversifying selection may not significantly confound the phylogenetic signal within the *S. aureus* genome in general.

Here we expand on these observations using *S. aureus* as a model population and a range of unlinked loci from all functional classes. The use of *S. aureus* has several advantages, as follows. (i) Extensive information on the population structure of this species is available through

Abbreviations: CAI, codon adaptation index; CE, cell envelope and cellular processes; FCT, fit to the consensus tree; IP, informational pathway; HK, housekeeping; MLST, multilocus sequence typing; MRSA, methicillin-resistant *S. aureus*; MSSA, methicillin-sensitive *S. aureus*; OR, orphan; UF, unknown function.

The GenBank/EMBL/DDBJ accession numbers for the sequences reported in this paper are DQ413277–DQ414234.

Two supplementary tables and two supplementary figures are available with the online version of this paper.

the generation of MLST data. (ii) Although recombination does occur (Robinson & Enright, 2004), *S. aureus* is basically clonal, which allows the reconstruction of a reasonably robust tree. This then facilitates comparisons between individual gene trees and a consensus tree. (iii) The data will provide a valuable phylogenetic framework for this important human pathogen.

We present sequence data from 33 unlinked gene loci representing a range of functions for 30 diverse *S. aureus* isolates. Supplementing these sequences with existing MLST data we reconstruct a phylogeny based on ~17.8 kb of concatenated sequence and compare each individual gene tree against a consensus phylogeny. We note no strong evidence that gene function, d_S/d_N ratio, G+C content or codon bias are strong predictors of phylogenetic reliability. This analysis does, however, provide a convenient rule of thumb that candidate phylogenetic markers should possess at least the average degree of sequence divergence (expressed as mean pairwise diversity, π) for all genes in the genome.

METHODS

Bacterial strains. We used a total of 30 *S. aureus* strains, 27 methicillin (formerly methicillin)-sensitive *S. aureus* (MSSA) sampled from cases of asymptomatic carriage ($n=9$), community-acquired disease ($n=5$) and hospital-acquired disease ($n=13$) recovered from Oxfordshire, UK. All these strains had previously been characterized by MLST, and were chosen to represent a diverse range of genotypes. A small number of duplicate STs were also included. We also included three strains from epidemic methicillin-resistant (MRSA) clones (EMRSA-3, EMRSA-4 and EMRSA-9) from global sources kindly donated by Dr Mark Enright, Department of Infectious Disease Epidemiology, Imperial College London, UK. See supplementary Table S1, available with the online version of this paper, for details of the strains.

Gene loci. We supplemented the MLST data already available for this strain collection (based on seven housekeeping genes) with a further 33 gene loci representing various functional categories to give a total dataset encompassing 40 loci, including 16S rRNA. These loci represent a range of functions, and are widely distributed across the chromosome (Fig. 1, Table 1). Genes were grouped into three functional classes, following Kuroda *et al.* (2001), adopted from the study of Kunst *et al.* (1997): informational pathways (IP;

DNA replication and processing, regulators; $n=9$), housekeeping (HK; central and intermediary metabolism; $n=13$), and cell envelope and cellular processes (CE; $n=5$). We also characterized conserved genes of unknown function (UF; $n=7$) and orphans (OR; unknown function, no similarity to other genes in the database; $n=6$). Genes of unknown function are referred to throughout using the SA ORF numbers proposed by Kuroda *et al.* (2001), except SA2439, which has subsequently been renamed *sasF* (Robinson & Enright, 2004).

DNA extraction, PCR and sequencing. DNA was purified using DNeasy kits (Qiagen) following the manufacturer's instructions. PCR was performed with an initial denaturation step of 3 min at 95 °C followed by 34 cycles of 30 s denaturation at 95 °C, 1 min annealing and 1 min extension at 72 °C. There was also a final extension step at 72 °C for 10 min. PCRs were successful for all genes in all strains except SA1621 (in strains H295 and H116) and SA0272 (in strain D22). As it was not possible to amplify these genes in all strains (presumably because of their absence), these genes were not included in the phylogenetic analysis. All genes were sequenced directly from purified PCR products using an ABI Prism 3700 sequencer. Primer sequences and annealing temperatures are given in supplementary Table S2. All sequences have been deposited at GenBank (accession numbers DQ413277–DQ414234).

Computation of sequence parameters. d_S/d_N ratios were calculated using the method of Nei & Gojobori (1986) as implemented in MEGA version 3.1 (Kumar *et al.*, 2004). Nucleotide diversity (π ; the mean percentage of polymorphic sites over all pairwise comparisons) and G+C content were calculated using MEGA version 3.1. The codon adaptation index (CAI) (Sharp & Li, 1987) was calculated by reference to the codon usage in ribosomal proteins using EMBOSS (Rice *et al.*, 2000).

Phylogenetic analysis. Of the 33 gene sequences generated, 30 were used for the phylogenetic analysis (two genes were not present in all strains, and the 16S rRNA fragment was invariant). These 30 genes were supplemented with the existing MLST data and a consensus Bayesian phylogeny was reconstructed from the concatenated sequences of all 37 genes representing 17 814 bp using MrBayes version 3.1 (Huelsenbeck & Ronquist, 2001; Ronquist & Huelsenbeck, 2003). This procedure uses a simulation technique, Markov chain Monte Carlo (MCMC), to approximate the posterior probabilities of alternative trees conditioned on the input data. As well as being very computationally efficient, the approach enables the sampling of a wide range of 'tree-space' rather than just locally optimum trees as in hill-climbing algorithms (for more details see <http://mr bayes.csit.fsu.edu/manual.php>). Four MCMC chains were run for 1 000 000 generations. The optimal trees were sampled every 100 generations

Fig. 1. Distribution of selected loci representing different functional categories around the *S. aureus* chromosome. Genes shown inside the ring are coded on the lagging strand; those on the outside are coded on the leading strand.

Table 1. Details of selected genes

All genes except the two indicated in the footnotes were found to be present in all strains.

Category	Gene	Position*	Fragment size (bp)	Function
CE	<i>aapA</i>	1732548	423	D-Serine/D-alanine/glycine transporter
CE	<i>pbpB</i>	1486656	474	Bifunctional type A penicillin-binding protein (PBP2)
CE	SA0272†	327449	450	Hypothetical protein similar to transmembrane protein Tmp7
CE	SA0817	920455	495	Hypothetical protein, similar to NADH-dependent flavin oxidoreductase
CE	<i>vicK</i>	25648	384	Two-component sensor histidine kinase
HK	<i>adhE</i>	164457	432	Alcohol-acetaldehyde dehydrogenase
HK	<i>arcC</i>	2723049	456	Carbamate kinase
HK	<i>aroE</i>	1629141	456	Shikimate dehydrogenase
HK	<i>glpF</i>	1296691	465	Glycerol kinase
HK	<i>gmk</i>	1191032	429	Guanylate kinase
HK	<i>hemH</i>	1886217	819	Ferrochetalase homologue
HK	<i>hutH</i>	10879	429	Histidine ammonia lyase
HK	<i>hutI</i>	2386381	807	Imidazalonepropionase
HK	<i>leuB</i>	2104034	849	3-Isopropylmalate dehydrogenase
HK	<i>pta</i>	1770243	474	Phosphate acetyltransferase
HK	SA0224	270714	456	Similar to 3-hydroxyacyl-CoA dehydrogenase
HK	<i>tpi</i>	835146	402	Triosephosphate isomerase
HK	<i>yqiL</i>	835146	516	Acetyl-CoA acetyltransferase
IP	<i>agrC</i>	2080353	390	Accessory gene regulator C
IP	<i>dnaC</i>	20770	414	Replicative DNA helicase
IP	<i>hsdR</i>	216859	399	Probable type 1 restriction enzyme restriction chain
IP	<i>luxS</i>	2186360	384	Autoinducer 2 production protein LuxS
IP	<i>sarA</i>	666721	294	Staphylococcal accessory regulator A
IP	<i>serS</i>	12793	453	Seryl-tRNA synthetase
IP	<i>sigB</i>	2118920	441	Sigma factor B
IP	<i>tufA</i>	590790	462	Translational elongation factor TU
OR	SA0139	158837	426	Hypothetical protein
OR	SA0268	324010	471	Hypothetical protein
OR	SA0740	847031	456	Hypothetical protein
OR	SA1619	1853499	417	Hypothetical protein
OR	SA1621‡	1854608	456	Hypothetical protein
OR	SA2445	2753901	459	Hypothetical protein
OTHER	SA0117	135490	435	Similar to rhizobactin siderophore biosynthesis protein
OTHER	SArRNA16	2234298	470	16S rRNA
UF	SA0013	18328	435	Conserved hypothetical protein
UF	SA0100	115153	444	Conserved hypothetical protein
UF	SA0275	331163	450	Conserved hypothetical protein
UF	SA0775	880970	405	Conserved hypothetical protein
UF	SA0778	884238	456	Conserved hypothetical protein
UF	SA1544	1764642	495	Hypothetical protein similar to soluble hydrogenase 42kD subunit
UF	<i>sasF</i>	2744355	432	Conserved hypothetical protein

*With respect to genome of N315 (Kuroda *et al.*, 2001).

†Absent in D22.

‡Absent in H295 and H116.

(with the first 2000 trees discarded as 'burn-in'). A 50 % majority rule consensus tree was then calculated using PAUP* version 4.0b10 (Swofford, 2000) with the posterior probabilities indicating the percentage of optimal trees supporting each node.

Fit to the consensus tree. As very variable genes make a larger contribution, in terms of informative sites, to the consensus tree

than very uniform genes, variable genes are more likely to show a closer fit. In order to draw independent comparisons between individual gene trees and the consensus, we constructed a further 37 consensus trees, in each case excluding a single gene. We then compared each of these consensus trees in turn with the gene tree corresponding to the excluded gene. We used the Shimodaira-Hasegawa (S-H) test (Shimodaira, 2002) in order to rank each gene

with respect to the differences in likelihood values between individual gene trees and the corresponding consensus tree (using the concatenated data as the reference). The S-H test was implemented in PAUP* version 4.0b10 (Swofford, 2000); a lower likelihood difference (S-H score) reflects a closer fit to consensus tree (FCT).

RESULTS

Table 2 gives the mean pairwise percentage nucleotide diversity (π), the mol% G + C content, the codon adaptation index (CAI) and the d_s/d_N ratios of all the gene loci

employed in this study. The genes are ranked according to the likelihood differences between individual gene trees and the consensus tree (FCT). The value of π for all genes was 1.28 %. Five of the six most uniform genes were classified as IP genes (16S rRNA, 0.0 %; *sarA*, 0.02 %; *tufA*, 0.2 %; *serS*, 0.3 %; and *sigB*, 0.4 %). At the other extreme, three genes appeared unusually diverse [*agrC* (IP), 5.5 %; *aapA* (CE), 5.3 %; and SA1619 (OR), 4.0 %]. Although the d_s/d_N ratio varies substantially both within and between gene classes, none of the genes showed evidence of positive selection. The orphans tended to exhibit low d_s/d_N ratios (mean 3.1,

Table 2. Sequence parameters for selected genes

Gene	Category	π	Mol% G + C	CAI	d_s/d_N	S-H score	FCT rank
SA2439	UF	0.017402	32.6	0.624	1.1	1345.368	1
<i>pbpB</i>	CE	0.009915	39.6	0.665	9	1565.546	2
SA1619	OR	0.040238	33.1	0.65	2.5	1708.409	3
<i>leuB</i>	HK	0.00953	35.8	0.537	8.7	1776.105	4
SA0740	OR	0.012154	29.5	0.593	3	1800.583	5
SA0775	UF	0.006123	34.1	0.689	∞	2056.379	6
<i>hemH</i>	HK	0.008434	34	0.645	13	2116.149	7
SA1544	UF	0.008046	33.4	0.595	7.3	2255.252	8
SA0224	HK	0.011473	38.8	0.533	5.8	2340.697	9
<i>luxS</i>	IP	0.008435	33.2	0.673	29	2507.735	10
SA0817	CE	0.014364	37.6	0.57	7.6	2536.186	11
SA2445	OR	0.018211	32.4	0.544	5	2573.84	12
<i>vicK</i>	CE	0.009325	36.5	0.527	∞	2604.608	13
<i>hutI</i>	HK	0.014635	37.2	0.532	23.5	2627.56	14
<i>aapA</i>	CE	0.05333	34.5	0.573	44.3	2671.307	15
<i>aroE</i>	HK (MLST)	0.010952	30.2	0.602	6.8	2677.064	16
<i>agrC</i>	IP	0.055196	31.5	0.568	13.5	2838.84	17
<i>sigB</i>	IP	0.003951	35.1	0.515	13	3011.601	18
<i>tpi</i>	HK (MLST)	0.010412	37.6	0.794	6.3	3043.117	19
<i>dnaC</i>	IP	0.013988	40.5	0.528	125	3080.7	20
SA0100	UF	0.00974	33.3	0.574	∞	3109.434	21
<i>pta</i>	HK (MLST)	0.006538	36.2	0.579	8.3	3151.953	22
SA0139	OR	0.018839	38.4	0.533	2.5	3220.862	23
SA0268	OR	0.007894	31.4	0.507	2.3	3271.318	24
SA0778	UF	0.002405	32.7	0.726	∞	3364.755	25
SA0275	UF	0.022545	29.6	0.628	27	3388.939	26
<i>yqiL</i>	HK (MLST)	0.007807	37.9	0.606	10.3	3436.345	27
SA0013	UF	0.013904	39.4	0.575	∞	3458.439	28
<i>hsdR</i>	IP	0.007711	29.9	0.538	11.5	3757.817	29
<i>gmk</i>	HK (MLST)	0.008309	33.5	0.633	11.7	4010.359	30
<i>glpF</i>	HK (MLST)	0.004106	40.8	0.648	15	4307.895	31
<i>arcC</i>	HK (MLST)	0.007888	38.5	0.539	6.3	4459.983	32
<i>serS</i>	IP	0.00339	34.2	0.737	∞	4661.175	33
<i>adhE</i>	HK	0.004256	40.3	0.6	12	4767.796	34
<i>hutH</i>	HK	0.004164	38	0.543	11	5480.02	35
<i>tufA</i>	IP	0.00198	38.5	0.891	5	5727.689	36
<i>sarA</i>	IP	0.0002	26.5	0.671	∞	6564.651	37
SA0272	CE	0.0144	30.9	0.604	6.5	—	—
SA1621	OR	0.031	34	0.521	4.1	—	—
16SrRNA	IP	0	50.2	—	—	—	—

median 2·8), suggesting a low level of functional constraint and rapid evolution. This is consistent with the non-essentiality of these genes, as indicated by their absence from the sequenced genomes of the closely related species *S. epidermidis*.

The phylogeny of *S. aureus* lineages

Data for 37 loci were concatenated to produce a total of ~17·8 kb for each of the 30 strains, and used to produce the unrooted Bayesian tree presented in Fig. 2. This tree is broadly consistent with one previously published based on the concatenated sequences of the seven MLST genes, but is more robust and contains no unresolved branches. The tree confirms the division into two main groups, as reported previously (Feil *et al.*, 2003; Holden *et al.*, 2004; Robinson *et al.*, 2005) with Group 1 being further subdivided into Groups 1a and 1b. ST55 is an exceptional genotype, previously being classified as Group 1 but appearing to fall at an intermediate position between the two main groups from the current data.

Group 1a contains the major MRSA clones ST36 (EMRSA-15), ST22 (EMRSA-16) and ST45 (the Berlin clone) (Aires de Sousa & de Lencastre, 2004; Oliveira *et al.*, 2002), as well as the common MSSA clone ST30 from which ST36 is thought to have evolved (Enright *et al.*, 2000). Interestingly, and in contrast to the MLST tree, these data suggest that ST45 and ST30 share a common ancestor. Group 1b contains no major nosocomial lineages; although ST59 was

found to be relatively common amongst intravenous drug users in Brighton, UK (Monk *et al.*, 2004), and is an important community-acquired MRSA from the USA (Pan *et al.*, 2005; Vandenesch *et al.*, 2003). Group 2 contains the related CC8 clones (EMRSA-1, 2, 4, 5, 6, 11 and 17), which includes the first MRSA lineage to be described (Crisostomo *et al.*, 2001), ST5 (EMRSA-3; the New York/Japan clone) (Oliveira *et al.*, 2002) and ST1, which is the genotype of sequenced strains MSSA476 (Holden *et al.*, 2004) and MW2 (Kuroda *et al.*, 2001). Group 2 also contains a relatively high number of sporadic or asymptomatic genotypes (e.g. ST20, ST9, ST13, ST101, ST7, ST97) and exhibits shorter branch lengths and lower clade credibility values than Group 1a.

Relationship between gene function and fit to the consensus tree

We ranked each gene tree with respect to its fit to a consensus tree (FCT) reconstructed excluding the gene under examination using the S-H test, as described in Methods (Table 2). All the genes showed significantly lower likelihood scores ($P < 0\cdot001$) against the consensus tree (compared with the concatenated data) using the S-H test. The gene showing the closest FCT (i.e. the smallest likelihood difference) was *sasF* (SA2439) which is of unknown function but likely to encode a surface-associated protein as it contains an LPXTG motif (Roche *et al.*, 2003); it was one of several putative cell-wall-associated genes used for a fine-scale study of the micro-evolution of MRSA clonal lineages by Robinson & Enright (2003). This result is surprising, as

Fig. 2. Bayesian reconstruction of *S. aureus* phylogeny based on the concatenated sequences of 37 gene fragments (~17·8 kb). The three subgroups are highlighted; it is unclear if ST55 should be assigned as a Group 2 genotype. The posterior probability scores are given on internal branches.

cell-wall-associated genes might be expected to be subject to diversifying selection pressure from the host immune response, and hence are likely candidates for frequent recombination. *sasF* exhibits reasonably high nucleotide diversity ($\pi = 1.7\%$) and the lowest d_S/d_N ratio of all the genes examined (1.1). The high degree of congruence of this gene to the consensus tree suggests that diversifying selection has not compromised the phylogenetic signal of the gene. The next highest scoring gene was *pbpB*, which encodes the bifunctional protein PBP2 (Pinho *et al.*, 2001). Although fulfilling an essential housekeeping function, PBP2 is an important target for β -lactam resistance (Leski & Tomasz, 2005) and vancomycin-intermediate glycopeptide resistance (Sieradzki & Tomasz, 1999). The third highest scoring gene in the S-H analysis is SA1619. This is an orphan of unknown function, although clearly one which has a stable and long-term association with *S. aureus* and is not prone to frequent transfer. Thus there are reasons for which each of the three top-scoring genes might have been avoided under classical MLST criteria.

With the exception of the very uniform informational genes, with their poor fit to the consensus tree, there is no obvious relationship between gene function and FCT. The housekeeping genes rank between 4th and 35th (Table 2) and in general score no better than cellular envelope genes or ORFans. It is noteworthy that three of the MLST genes, *gmk*, *glpF* and *arcC*, rank 30th–32nd respectively and only out-rank those genes which are extremely uniform. This analysis also confirms a previous suggestion (Feil *et al.*, 2003) that *arcC* in particular possesses an atypical phylogenetic signal.

Relationship between nucleotide diversity and fit to the consensus tree

To examine the role of other sequence parameters we plotted the FCT of each gene against π , G + C content, d_S/d_N ratio and codon bias. Owing to very low levels of diversity, *sarA* was excluded from these analyses. Plotting π against FCT confirms that more diverse genes tend to show a closer fit to the consensus (linear plot: $R^2 = 0.111$, $P = 0.047$; Fig. 3a; Spearman's rank correlation coefficient = -0.508 , $P = 0.002$). The use of a quadratic plot increases the R^2 to 0.329 ($P = 0.001$; not shown), suggesting that the relationship between π and FCT is not linear. If the pairwise diversities are ranked, which provides a closer fit of the residuals to a normal distribution (by controlling for the effect of extreme values), a quadratic plot gives an R^2 of 0.441 ($P < 0.0001$; Fig. 3b). This plot demonstrates that the relationship between diversity and FCT only holds for the more uniform genes.

To examine this further we divided the genes into two equal groups according to π and plotted the rank in diversity against FCT as a linear trend for each group. Examining the most uniform 18 genes separately reveals a linear correlation of increasing FCT with increasing π ($R^2 = 0.342$, $P = 0.011$), whereas the 18 most diverse genes did not reveal a significant trend ($R^2 = 0.024$, $P = 0.54$; plots not shown). Thus for

Fig. 3. Relationship between mean pairwise nucleotide diversity (π) and FCT (SH score). A lower SH score reflects a smaller difference in likelihood score between the gene and the consensus trees, thus a higher FCT. (a) Plotted using the values of π with a linear regression line. (b) Ranks of π are plotted and the regression line is quadratic; this clearly illustrates that the linear relationship between π and FCT only holds for low values of π (i.e. more uniform genes).

genes which fall below a threshold level of π (in this case approximately 1%), pairwise nucleotide diversity is a strong predictor of phylogenetic reliability. For genes above 1% diversity there is no obvious relationship, but the observation that two of the three very diverse genes show a modest FCT (Fig. 3a) suggests that there is also an upper threshold of π with respect to FCT. The most diverse gene, *agrC*, is ranked 17th in terms of FCT, which confirms the discrepancies between *agr* groups and *S. aureus* phylogeny discussed elsewhere (Robinson *et al.*, 2005).

We also examined the correlation between FCT and G + C content, d_S/d_N ratio and codon bias. We noted no evidence of a correlation with d_S/d_N ratio or codon bias (data not shown), but a weak correlation with G + C content ($R^2 = 12.7\%$, $P = 0.033$; see supplementary Fig. S1). The significance of this association with G + C content is unclear and requires further analysis.

DISCUSSION

We have presented an intra-species tree for *S. aureus* based on ~ 17.8 kb of concatenated sequence which provides hypotheses concerning the relatedness between the major

MRSA lineages. Although this is an improvement on the existing tree, the branching order cannot be reconstructed with complete confidence in some parts of the tree. Would the tree be improved by the addition of yet more data? Rokas *et al.* (2003) examined phylogenetic congruence in eight yeast species and concluded that the concatenated data of a minimum of 20 genes are required to produce a robust tree. Although in terms of nucleotide sites our dataset of 38 gene fragments is of a similar size to the 20 genes of Rokas *et al.* (2003), the use of a higher number of independently evolving genes should increase the performance of the data. Therefore we feel that the intra-species tree we present here would not be greatly improved by the addition of yet more data. The broad consistency of this phylogeny with the basic groupings previously inferred from MLST genes (Feil *et al.*, 2003), *sas* genes (Robinson *et al.*, 2005), adhesin genes (Kuhn *et al.*, 2006), AFLP clustering (Melles *et al.*, 2004), PFGE (Grundmann *et al.*, 2002) and microarray analysis (Lindsay *et al.*, 2006) provides additional support.

The topology within Group 2 remains relatively poorly supported, and contrasts with the much longer branches evident in Group 1a. This difference between these groups has also been noted in an analysis of MLST and *sas* genes (Robinson *et al.*, 2005). One possibility is that the globally disseminated Group 1a clones (clonal complexes – ‘CCs’ – 30, 45 and 22) may be particularly efficient at out-competing close relatives and that the longer branch lengths in Group 1a reflect a higher rate of stochastic extinction than in Group 2. The relatively poor clade credibility scores in Group 2 are also consistent with a higher rate of recombination in Group 2 strains, although comparisons of the two groups using various tests for recombination did not produce strong evidence to support this view (data not shown).

Our data also suggest that Group 1 strains can be further subdivided into Group 1a and Group 1b, a division not recognized in previous phylogenetic studies. Although there is generally little association between phylogenetic distribution and epidemiological source, it is noteworthy that Group 1b contains no major nosocomial lineages, whereas Group 1a contains the two major MRSA clones currently circulating in the UK (STs 36 and 22), as well as the Berlin clone (ST45). Future studies aimed at identifying the genetic factors underlying the ability to rapidly disseminate might therefore focus on comparisons between Group 1a and Group 1b strains. An interesting observation in this context is the high degree of divergence between Group 1a and Groups 1b and 2 at *aapA* (see supplementary Fig. S2); an examination of the region surrounding this gene might therefore shed some light on the epidemiological differences between the two groups.

Although the phylogenetic emphasis of this study was on the relationships between the major clonal lineages, we included duplicates of four STs (5, 22, 36 and 121). In each case, these duplicates differed at five or fewer positions in the concatenated sequence of 17 814 sites (<0.0004%). This is consistent with a comparative genome analysis of two ST1

isolates (MSSA476 and MW2) which only revealed 285 single base changes in all orthologous gene pairs (~1 in 10,000 sites) (Holden *et al.*, 2004). These results confirm the high degree of genetic relatedness between isolates sharing identical STs. However, a more extensive investigation of intra-clonal differences has proved successful in providing detailed hypotheses concerning the emergence of closely related MRSA clones (Robinson & Enright, 2003). This study utilized the highly variable *sas* genes, and our current results suggest that these genes might also be highly informative for reconstructing deeper relationships within the *S. aureus* population. A second study, utilizing variable adhesin genes, provided some evidence that recombination is more common within, rather than between, clonal complexes (Kuhn *et al.*, 2006).

Gene function, diversity and informative trees

These data are not only relevant for studies on *S. aureus*, but also provide clues as to the extent to which the current criteria for choosing gene loci for phylogenetic, systematics or epidemiological studies can be justified or relaxed. Here we find little evidence to justify the current emphasis on housekeeping genes, at least on an intra-species level, and indeed our results for *S. aureus* suggest that the MLST genes for this species rate amongst the poorest phylogenetic markers. In contrast, the three genes which score highly against the consensus tree are putatively associated with the cell wall (*sasF*), modified in antibiotic-resistant strains (*pbpB*) or an orphan (SA1619), all of which would have been avoided under classical MLST criteria.

We suggest that the emphasis on gene choice for intra-species phylogenetic markers should be shifted to the more tangible parameter of nucleotide diversity, with gene function being regarded as secondary. Clearly, gene function and diversity are not always independent; ‘informational pathway’ genes (in particular 16S rRNA) should generally be avoided due to the extremely low levels of diversity of these genes. It is not clear what determines the point at which extra variation ceases to improve the tree, and more specifically why variation in reasonable excess of 1% generally does not result in a closer fit to the consensus phylogeny. Nevertheless, this analysis provides a convenient ‘rule of thumb’ for identifying genes which are likely to contain sufficient diversity, i.e. those containing at least the average for all genes. Our results also raise the possibility of a correlation between G + C content and closeness of fit to a consensus tree. Given the large number of potential candidate loci for each gene it may therefore also be sensible to avoid those with extreme G + C contents.

We emphasize that we do not advocate changes to any established MLST scheme. The current MLST scheme for *S. aureus* has proved extremely successful in understanding the population structure of this species and for assigning isolates to particular lineages. In a highly clonal organism, almost any gene will typically provide the same basic lineage assignments – in the case of *S. aureus* this is clear from the

broad consistency of different genes as well as pan-genome techniques such as PFGE (Grundmann *et al.*, 2002) and microarray analysis (Lindsay *et al.*, 2006). However, individual genes may vary in their utility to reconstruct the relationships between these lineages, and we find no evidence to suggest that MLST genes can be considered the most reliable in this regard.

Concluding remarks

We present the most robust tree to date of the natural *S. aureus* population, and identify three distinct groups within the population. We propose an emphasis on gene diversity, rather than gene function, when identifying suitable phylogenetic markers. Although this may necessitate preliminary work on candidate loci before final genes are chosen, we argue that this represents a sensible investment of resources. Finally, our analysis differs from studies on more deep-rooted phylogenies (i.e. those between genera or orders) (Zeigler, 2003). In this case, the presence of sufficient diversity is not likely to be problematic and the use of 'core' genes may well be justified. At an intra-species level, however, given the choice of many candidate ubiquitous genes, we argue that the presence of sufficient diversity should be considered first and foremost, and other considerations relating to gene function should be secondary.

ACKNOWLEDGEMENTS

This work was funded by an MRC Career Development Award to E. J. F. We are grateful to Eduardo Rocha for calculation of the CAI values, to Mark Enright for the provision of strains and to Ashley Robinson for constructive comments on the manuscript.

REFERENCES

- Aires de Sousa, M. & de Lencastre, H. (2004). Bridges from hospitals to the laboratory: genetic portraits of methicillin-resistant *Staphylococcus aureus* clones. *FEMS Immunol Med Microbiol* **40**, 101–111.
- Baptiste, E., Susko, E., Leigh, J., MacLeod, D., Charlebois, R. L. & Doolittle, W. F. (2005). Do orthologous gene phylogenies really support tree-thinking? *BMC Evol Biol* **5**, 33.
- Crisostomo, M. I., Westh, H., Tomasz, A., Chung, M., Oliveira, D. C. & de Lencastre, H. (2001). The evolution of methicillin resistance in *Staphylococcus aureus*: similarity of genetic backgrounds in historically early methicillin-susceptible and -resistant isolates and contemporary epidemic clones. *Proc Natl Acad Sci U S A* **98**, 9865–9870.
- 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.
- Feil, E. J., Maiden, M. C., Achtman, M. & Spratt, B. G. (1999). The relative contributions of recombination and mutation to the divergence of clones of *Neisseria meningitidis*. *Mol Biol Evol* **16**, 1496–1502.
- Feil, E. J., Smith, J. M., Enright, M. C. & Spratt, B. G. (2000). Estimating recombinational parameters in *Streptococcus pneumoniae* from multilocus sequence typing data. *Genetics* **154**, 1439–1450.
- Feil, E. J., Cooper, J. E., Grundmann, H. & 9 other authors (2003). How clonal is *Staphylococcus aureus*? *J Bacteriol* **185**, 3307–3316.
- Gevers, D., Cohan, F. M., Lawrence, J. G. & 8 other authors (2005). Opinion: re-evaluating prokaryotic species. *Nat Rev Microbiol* **3**, 733–739.
- Grundmann, H., Hori, S., Enright, M. C., Webster, C., Tami, A., Feil, E. J. & Pitt, T. (2002). Determining the genetic structure of the natural population of *Staphylococcus aureus*: a comparison of multi-locus sequence typing with pulsed-field gel electrophoresis, randomly amplified polymorphic DNA analysis, and phage typing. *J Clin Microbiol* **40**, 4544–4546.
- Hanage, W. P., Fraser, C. & Spratt, B. G. (2005). Fuzzy species among recombinogenic bacteria. *BMC Biol* **3**, 6.
- 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.
- Huelsenbeck, J. P. & Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754–755.
- Jolley, K. A., Kalmusova, J., Feil, E. J., Gupta, S., Musilek, M., Kriz, P. & Maiden, M. C. (2000). Carried meningococci in the Czech Republic: a diverse recombining population. *J Clin Microbiol* **38**, 4492–4498.
- Kuhn, G., Francioli, P. & Blanc, D. S. (2006). Evidence for clonal evolution among highly polymorphic genes in methicillin-resistant *Staphylococcus aureus*. *J Bacteriol* **188**, 169–178.
- Kumar, S., Tamura, K. & Nei, M. (2004). MEGA3: integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Brief Bioinform* **5**, 150–163.
- Kunst, F., Ogasawara, N., Moszer, I. & 148 other authors (1997). The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature* **390**, 249–256.
- Kuroda, M., Ohta, T., Uchiyama, I. & 34 other authors (2001). Whole genome sequencing of methicillin-resistant *Staphylococcus aureus*. *Lancet* **357**, 1225–1240.
- Leski, T. A. & Tomasz, A. (2005). Role of penicillin-binding protein 2 (PBP2) in the antibiotic susceptibility and cell wall cross-linking of *Staphylococcus aureus*: evidence for the cooperative functioning of PBP2, PBP4, and PBP2A. *J Bacteriol* **187**, 1815–1824.
- Lindsay, J. A., Moore, C. E., Day, N. P., Peacock, S. J., Witney, A. A., Stabler, R. A., Husain, S. E., Butcher, P. D. & Hinds, J. (2006). Microarrays reveal that each of the ten dominant lineages of *Staphylococcus aureus* has a unique combination of surface-associated and regulatory genes. *J Bacteriol* **188**, 669–676.
- Maiden, M. C., Bygraves, J. A., Feil, E. & 10 other authors (1998). Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci U S A* **95**, 3140–3145.
- Melles, D. C., Gorkink, R. F., Boelens, H. A. & 8 other authors (2004). Natural population dynamics and expansion of pathogenic clones of *Staphylococcus aureus*. *J Clin Invest* **114**, 1732–1740.
- Monk, A. B., Curtis, S., Paul, J. & Enright, M. C. (2004). Genetic analysis of *Staphylococcus aureus* from intravenous drug user lesions. *J Med Microbiol* **53**, 223–227.
- Nei, M. & Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* **3**, 418–426.
- Oliveira, D. C., Tomasz, A. & de Lencastre, H. (2002). Secrets of success of a human pathogen: molecular evolution of pandemic clones of methicillin-resistant *Staphylococcus aureus*. *Lancet Infect Dis* **2**, 180–189.

- Pan, E. S., Diep, B. A., Charlebois, E. D., Auerswald, C., Carleton, H. A., Sensabaugh, G. F. & Perdreau-Remington, F. (2005). Population dynamics of nasal strains of methicillin-resistant *Staphylococcus aureus* – and their relation to community-associated disease activity. *J Infect Dis* **192**, 811–818.
- Pinho, M. G., Filipe, S. R., de Lencastre, H. & Tomasz, A. (2001). Complementation of the essential peptidoglycan transpeptidase function of penicillin-binding protein 2 (PBP2) by the drug resistance protein PBP2A in *Staphylococcus aureus*. *J Bacteriol* **183**, 6525–6531.
- Rice, P., Longden, I. & Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16**, 276–277.
- Robinson, D. A. & Enright, M. C. (2003). Evolutionary models of the emergence of methicillin-resistant *Staphylococcus aureus*. *Antimicrob Agents Chemother* **47**, 3926–3934.
- Robinson, D. A. & Enright, M. C. (2004). Evolution of *Staphylococcus aureus* by large chromosomal replacements. *J Bacteriol* **186**, 1060–1064.
- Robinson, D. A., Monk, A. B., Cooper, J. E., Feil, E. J. & Enright, M. C. (2005). Evolutionary genetics of the accessory gene regulator (*agr*) locus in *Staphylococcus aureus*. *J Bacteriol* **187**, 8312–8321.
- Roche, F. M., Massey, R., Peacock, S. J., Day, N. P., Visai, L., Speziale, P., Lam, A., Pallen, M. & Foster, T. J. (2003). Characterization of novel LPXTG-containing proteins of *Staphylococcus aureus* identified from genome sequences. *Microbiology* **149**, 643–654.
- Rokas, A., Williams, B. L., King, N. & Carroll, S. B. (2003). Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* **425**, 798–804.
- Ronquist, F. & Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572–1574.
- Sharp, P. M. & Li, W. H. (1987). The codon adaptation index – a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* **15**, 1281–1295.
- Shimodaira, H. (2002). An approximately unbiased test of phylogenetic tree selection. *Syst Biol* **51**, 492–508.
- Sieradzki, K. & Tomasz, A. (1999). Gradual alterations in cell wall structure and metabolism in vancomycin-resistant mutants of *Staphylococcus aureus*. *J Bacteriol* **181**, 7566–7570.
- Spratt, B. G. & Maiden, M. C. (1999). Bacterial population genetics, evolution and epidemiology. *Philos Trans R Soc Lond B Biol Sci* **354**, 701–710.
- Swofford, D. L. (2000). PAUP* – *Phylogenetic Analysis Using Parsimony**, and Other Methods. Sunderland, MA: Sinauer Associates.
- Vandenesch, F., Naimi, T., Enright, M. C. & 8 other authors (2003). Community-acquired methicillin-resistant *Staphylococcus aureus* carrying Panton-Valentine leukocidin genes: worldwide emergence. *Emerg Infect Dis* **9**, 978–984.
- Zeigler, D. R. (2003). Gene sequences useful for predicting relatedness of whole genomes in bacteria. *Int J Syst Evol Microbiol* **53**, 1893–1900.