Evolution of Paralogous Genes: Reconstruction of Genome Rearrangements Through Comparison of Multiple Genomes Within *Staphylococcus aureus*

Takeshi Tsuru,†‡ Mikihiko Kawai,‡§ Yoko Mizutani-Ui,‡ Ikuo Uchiyama,‖ and Ichizo Kobayashi*†‡§*

*Department of Medical Genome Sciences, Graduate School of Frontier Science, University of Tokyo, Tokyo, Japan; †Graduate Program in Biophysics and Biochemistry, Graduate School of Science, University of Tokyo, Tokyo, Japan; ‡Division of Molecular Biology, Institute of Medical Science, University of Tokyo, Tokyo, Japan; §Division of Pathology, Immunology and Microbiology, Graduate School of Medicine, University of Tokyo, Tokyo, Japan; and ‖Laboratory of Genome Informatics, Natural Institute for Basic Biology, National Institutes of Natural Sciences, Okazaki, Japan

Analysis of evolution of paralogous genes in a genome is central to our understanding of genome evolution. Comparison of closely related bacterial genomes, which has provided clues as to how genome sequences evolve under natural conditions, would help in such an analysis. With species *Staphylococcus aureus*, whole-genome sequences have been decoded for seven strains. We compared their DNA sequences to detect large genome polymorphisms and to deduce mechanisms of genome rearrangements that have formed each of them. We first compared strains N315 and Mu50, which make one of the most closely related strain pairs, at the single-nucleotide resolution to catalogue all the middle-sized (more than 10 bp) to large genome polymorphisms such as indels and substitutions. These polymorphisms include two paralogous gene sets, one in a tandem paralogue gene cluster for toxins in a genomic island and the other in a ribosomal RNA operon. We also focused on two other tandem paralogue gene clusters and type I restriction-modification (RM) genes on the genomic islands. Then we reconstructed rearrangement events responsible for these polymorphisms, in the paralogous genes and the others, with reference to the other five genomes. For the tandem paralogue gene clusters, we were able to infer sequences for homologous recombination generating the change in the repeat number. These sequences were conserved among the repeated paralogous units likely because of their functional importance. The sequence specificity (S) subunit of type I RM systems showed recombination, likely at the homology of a conserved region, between the two variable regions for sequence specificity. We also noticed novel alleles in the ribosomal RNA operons and suggested a role for illegitimate recombination in their formation. These results revealed importance of recombination involving long conserved sequence in the evolution of paralogous genes in the genome.

Introduction

Various forms of genome rearrangements are important in the plasticity of genomes and contribute greatly to genome evolution. One of the specific issues relevant to genome evolution through rearrangement is evolution of multiple homologous genes present in a genome, often called paralogous genes. Earlier, it was addressed through comparison within a genome and between distantly related genomes. For several years, comparative analyses of closely related bacterial genome sequences have been providing numerous novel insights about genome evolution in general and paralogue evolution in particular (Alm et al. 1999; Rocha and Blanchard 2002; Rocha 2004).

Close genome comparison revealed various genome rearrangements involving repeated genes within a genome and prevalence of horizontal gene transfer throughout bacterial and archaeal worlds. For example, our laboratory has compared two genome sequences within species *Helicobacter pylori* and found a characteristic mode of insertion of restriction-modification (RM) gene complexes (Nobusato et al. 2000). We have been developing a tool, Comparative Genome Analysis Tool (CGAT), to compare two bacterial genome sequences to identify middle-sized to large genome polymorphisms and to help inference about their formation (Uchiyama, Higuchi, and Kobayashi 2000). Availability of as many as seven complete genome sequences within the species *Staphylococcus aureus* (Kuroda et al. 2001; Baba et al. 2002; Holden et al. 2004; Ohta et al. 2004; Gill et al. 2005) (http://www.genome.ou.edu.) provides a unique opportunity for such an analysis.

Staphylococcus aureus represents a gram-positive eubacterium with low GC content (Kuroda et al. 2001). Staphylococci are normal inhabitants on skin and mucous membranes of warm-blooded animals and can become pathogens. In addition, this bacterium has developed resistance to practically all types of antibiotics (Hiramatsu et al. 2001). The conserved synteny in the seven genomes helped in whole-genome comparison with each other, which revealed that a major diversity of *S. aureus* strains was associated with a variety of mobile elements: prophages, transposons, insertion sequences, and other genomic islands (Lindsay and Holden 2004; Gill et al. 2005). The intraspecific variety with respect to pathogenicity and antibiotic resistance was, at least partly, found associated with these mobile elements (Kuroda et al. 2001; Baba et al. 2002; Holden et al. 2004; Gill et al. 2005). A recent comparison between N315 and Mu50 genomes exhaustively inspected all the putative open reading frames (ORFs) for difference (Ohta et al. 2004).

In *S. aureus* only little is known about molecular mechanisms of gene transfer and genome rearrangements. Conjugation was hypothesized for apparent chromosomal replacements inferred from multilocus sequence typing (Robinson and Enright 2004). A class of genomic islands can move from one cell to another with the aid of a helper phage (Ruzin, Lindsay, and Novick 2001). Staphylococcal

Key words: genome comparison, genome rearrangements, deletion, recombination, restriction-modification.

E-mail: ikobaya@ims.u-tokyo.ac.jp.

Mol. Biol. Evol. 23(6):1269–1285. 2006
doi:10.1093/molbev/msk013
Advance Access publication April 6, 2006

cassette chromosomes and several types of genomic islands are believed to integrate into the chromosome by their own site-specific recombinase (Katayama, Ito, and Hiramatsu 2000; Ruzin, Lindsay, and Novick 2001).

The genome sequence analysis of N315 and Mu50 strains (Kuroda et al. 2001), in which this laboratory was involved, prompted us to study rearrangement mechanisms that have formed their differences. We focused on middle-sized (more than 10 bp) to large polymorphisms, which were located at the toxin paralogue gene cluster in genomic island νSaα, the ribosomal RNA operon, and other short repeats with variable configurations.

Two genomic islands, νSaα and νSaβ, identified in all the strains of *S. aureus* examined so far (Baba et al. 2002), carry three tandem paralogous gene clusters: staphylococcal superantigen-like (*ssl* or *set*) (Lina et al. 2004) gene cluster and lipoprotein (*lpl*) gene cluster in νSaα and serine protease (*spl*) gene cluster in νSaβ. For the *ssl* cluster in the seven strains, Fitzgerald et al. (2003) suggested that an ancestral strain with a full complement of *ssl* genes underwent multiple gene losses by some recombination mechanisms. Here we have considered molecular mechanisms of genome rearrangements in more detail for the *ssl* cluster as well as for the *lpl* cluster and the *spl* cluster.

Another feature of these genomic islands is the presence of type I RM genes linked with those tandem gene clusters. A type I RM system comprises three genes, *hsdR*, *hsdM*, and *hsdS*, which are tightly linked (Murray 2000) with interesting exceptions (Schouler et al. 1998; Rocha and Blanchard 2002). In the seven *S. aureus* genomes, only *hsdM* and *hsdS* are found both in νSaα and νSaβ. A homologue of *hsdR* (SA0189 for N315) has been identified at a locus distant from these two islands in all these genomes (Kuroda et al. 2001). Though the deduced amino acid sequences for *hsdM* are almost identical among all the 14 alleles, those for *hsdS* show divergence (Baba et al. 2002). We have characterized the variation found in the *hsdS* genes.

The ribosomal RNA (*rrn*) operon is involved in various genome rearrangements. Homologous recombination via extensive homology of rDNAs can cause genome-wide reorganization (Hill 1999) and homogenization of rDNAs and their flanking regions (Liao 2000). The 16S-23S rDNA intergenic spacer region is known to be polymorphic (Gurtler and Stanisich 1996). Several types of rearrangements there were reported and explained also in terms of homologous recombination (Harvey et al. 1988; Lan and Reeves 1998; Privitera et al. 1998; Gurtler 1999). For *S. aureus*, 10 different types of sequences of this region were reported (Gurtler and Barrie 1995; Gurtler and Mayall 2001). These analyses included two complete genomes and three unfinished genomes available at the time. We examined whole sets of *rrn* operons of the seven strains and considered how their differences were generated.

Materials and Methods

Sequence Sources

Annotated whole-genome sequences of the following *S. aureus* strains have been released: N315 and Mu50 (Kuroda et al. 2001; Ohta et al. 2004), MW2 (Baba et al. 2002), MRSA252 and MSSA476 (Holden et al. 2004), COL (Gill et al. 2005), and NCTC8325 (RefSeq: NC_007795). We downloaded them from GenBank (http://www.ncbi.nlm.nih.gov/Genbank/). All the other sequence data were from GenBank or REBASE (http://rebase.neb.com).

Software and Programs

BlastN or BlastP (http://www.ncbi.nlm.nih.gov/Blast/) for homology search and ClustalW (http://www.ebi.ac.uk/clustalw/) for sequence alignment were introduced with default parameters. The resulting alignments were followed by manual refinement on Se-Al (http://evolve.zoo.ox.ac.uk/software.html?id=seal) when necessary. The dot plot analyses for pairwise and multiple comparison were performed by DOTTUP and POLYDOT (http://emboss.sourceforge.net), respectively. Each of these software and programs were downloaded from the ftp site and run on Mac OS X (version 10.3.7).

Screening of Polymorphisms Between Strains N315 and Mu50

The genome-wide polymorphisms between N315 and Mu50 were screened by pairwise alignment on the CGAT (Uchiyama, Higuchi, and Kobayashi 2000). CGAT has organized the genome alignment by bidirectional best-hit analysis based on all-against-all comparison of BlastN for pairs of 2-kb segments with 200 bp each overlapping. All the polymorphic sites were identified by visual inspection as breaks within the whole-genome alignment in 500-bp windows of CGAT. This manual screening covered all the macroscopic differences of more than 10 bp.

Criteria for Classification of Putative Rearrangement Events

We tentatively classified recombination into four types: site-specific recombination, transposase-mediated recombination, homologous recombination, and illegitimate recombination.

Occurrence of homologous recombination and some sort of illegitimate recombination between similar DNA sequences was inferred from the presence of similar sequences at the sites of rearrangement in genome comparison. Dependence of homologous recombination frequency on the homology length varies among organisms and processes (Fujitani, Yamamoto, and Kobayashi 1995). In *Bacillus subtilis*, the closest relative of *S. aureus* so far reported with respect to this process, an apparent lower limit is reported to be 70 bp (Khasanov et al. 1992). Frequency of homologous recombination decreases very rapidly as the two sequences diverge (Datta et al. 1997; Vulic et al. 1997; Fujitani and Kobayashi 1999). From these data, we regarded the repeats of more than 80 bp long sharing as much as 90% nucleotide sequence identity in present data as a strong candidate for a remnant of homologous recombination.

Illegitimate recombination events can be classified into two: those between short repeats and those between sequences with no or very little similarity (Michel 1999). The former class can take place between sequences with

Table 1
Larger Genome Polymorphisms Between Strain N315 and Strain Mu50

ID[a]	Position (N315/Mu50)[b]	Description	Feature	Reference
#2	444496–445737/(470253–470254)	An indel within *ssl* tandem gene cluster	On a genomic island (νSaα)	Fitzgerald et al. (2003) and this study
#4	507789–507979/532103–532184	A substitution within 16S-23S rDNA intergenic spacer region of *rrn* operon	Interspersed repeats	Gurtler (1999) and this study
#6	777187–777242/(801435–801436)	An indel with respect to STAR	Genome-wide interspersed repeats	T. Tsuru, I. Uchiyama, and I. Kobayashi (unpublished data)
#7	850616–850777/(889293–889394)	An indel within serine-aspartate dipeptide repeat region of *clfA*	Interspersed repeats	McDevitt and Foster (1995), Foster and Hook (1998)
#13	2005322–2049592/2083116–2126182	Localized multiple polymorphisms in homologous prophages, ϕSa3	Prophage	Brussow, Canchaya, and Hardt (2004)
#16	(2475685–2475686)/2546705–2546895	An indel with respect to STAR	Genome-wide interspersed repeats	T. Tsuru, I. Uchiyama, and I. Kobayashi (unpublished data)

[a] The polymorphic sites of interest were numbered clockwise from the predicted replication origin.

[b] For indels, the region of the apparent insert on one genome and the base pairs adjacent to the apparent deletion on the other genome are presented with the latter base pairs in parentheses. For substitution, the nonhomologous region is presented for each. For homologous prophages, a region of prophage is presented for each.

much shorter homology than homologous recombination through such molecular events as simple slipped misalignment, sister chromosome slipped misalignment, or single-strand annealing (Lovett 2004). The latter one was reported to be caused by errors of DNA gyrase and topoisomerase I in *Escherichia coli* (Michel 1999). We expediently set the threshold between the two types of illegitimate recombination at 3 bp: equal or more than 3 bp versus less than 3 bp. In these mechanisms, there is strong negative dependence of the recombination frequency on the distance between the repeats (Chedin et al. 1994; Lovett et al. 1994), and it is estimated that less than hundreds of base pairs should be required for efficient processes due to their involvement with replication machinery (Michel 1999; Lovett 2004). Therefore, we assumed that the distance between the short recombining sequences has to be less than 1 kb in order to support an event of illegitimate recombination.

Larger Polymorphisms and Smaller Polymorphisms

The 19 polymorphic sites of interest, which exclude indels of mobile elements, were examined in CGAT for possible linkage with other mobile elements or genome-wide interspersed repeats (see *Results and Discussion*, *Macroscopic Polymorphisms Between N315 and Mu50*). The polymorphisms that are supposedly related to other mobile elements or genome-wide interspersed repeats were grouped as “larger polymorphisms” (see table 1, Feature). The others, the sequences of which were accordingly not homologous to the rest of the chromosome or other mobile elements, were designated as “smaller polymorphisms” (see table 2 below).

Analysis of Smaller Polymorphisms

We employed DOTTUP to arrange a dot plot matrix between N315 and Mu50 with appropriate parameters for each site of smaller polymorphisms. In cases of multiple repeats estimated as more than three copies by eye, Tandem Repeats Finder (http://tandem.bu.edu/trf/trf.html) was applied with default parameters in order to define each repeat unit and its copy number. The sequence identities of direct repeats neighboring indels were extracted manually from their optimized multiple alignments by ClustalW. Lengths of these direct repeat sequences were decided as the maximal lengths keeping 90% base pair identity or more.

The corresponding sites on the other five genomes were identified by BlastN search with N315 and Mu50 sequences as queries. Then the multiple dot plot comparison between the seven genomes was performed on POLYDOT.

Analysis of Paralogous Genes

For three tandem paralogue clusters within genomic islands, νSaα and νSaβ, the nucleotide sequences of the entire cluster and adjacent 200-bp regions were subjected to the dot plot analysis by POLYDOT with a 15-bp window.

For *ssl* and *spl* cluster, repeat sequences flanking each indel was sought by eye on the optimized multiple alignment of the three recombination joints in the relevant genome pairs with ClustalW for calculation of the length and sequence identity. N315-Mu50 strain pair gave the best alignment for the red indel in the *ssl* cluster, N315-MW2 gave the best for the orange indel in the *ssl* cluster, while N315-COL pair gave the best for the *spl* cluster.

For the *lpl* cluster and the *spl* cluster, we constructed Neighbor-Joining phylogenetic trees on the nucleotide *p*-distances (Nei and Kumar 2001) and then assigned each clustering group as to share 85% or more sequence identity. This phylogenetic grouping of the homologous ORFs was displayed in 12 and 7 distinct colors in figure 3*A* and *B* (see below) and figure 4*A* and *B* (see below), respectively.

For the *hsdS* genes of type I RM systems, the predicted amino acid sequences of those on νSaα and νSaβ in the sequenced strains and on their putative homologue on *etd* pathogenicity island in strain TY104 (Yamaguchi et al. 2002) were compared and it was confirmed that they are grouped into seven types (Baba et al. 2004) (see fig. 5*B* below, type). A multiple alignment was constructed by ClustalW for representatives of the above seven types

Table 2
Smaller Polymorphisms Between Strain N315 and Strain Mu50

Dot Plot Pattern in Figure 1[a]	ID[b]	Position (N315/Mu50)[c]	Repeat or Substitution Configuration (N315/Mu50) (bp)[a]	Flanking Direct Repeat		In the Other Five Strains[d] (N315 Type/Mu50 Type)
				Length (bp)	Identity (%)	
(*A*)	#3	485155–485543/509670–509857	188–13–188/188	188	91	5/0
	#9	1421765–1421775/1498094–1498170	11/11–55–11	11	100	0/5
	#12	1808424–18808554/1886338–1886349	12–107–12/12	12	100	5/0
	#14	2054307–2054473/2130897–2130906	10–147–10/10	10	100	5/0
	#15	2240390–2240619/2310110–2310119	10–210–10/10	10	90	5/0
	#17	2476933–2476988/2548143–2548168	26–4–26/26	26	92	5/0
(*B*)	#10	1423802/1500197–1500213	1/1–15–1	1	100	0/5
	#11	1684220–1684241/1760633–1760634	2–18–2/2	2	100	5/0
	#18	(2499013–2490014)[e]/2570194–2570207	0/14	NR	NR	0/5
(*C*)	#1	4315043507/4315143468	38–(2–38) × 8/38–(2–38) × 7	40	91	—[f]
	#5	526783–526833/550988–551080	9–(12–9) × 2/9–(12–9) × 4	21	100	—[g]
	#19	2560410–2562528/2631604–2632954	199–(185–199) × 5/199–(185–199) × 3	384	98	—[h]
(*D*)	#8	994370–994545/1070654–1070874	176/221	NR	NR	5/0

NOTE.—NR, not relevant.

[a] See figure 1.

[b] The polymorphic sites of interest were numbered clockwise from the predicted replication origin.

[c] Corresponds to a range of the polymorphic region depicted as a gray line in figure 1.

[d] The sequences of the locus were compared in the other five strains, NCTC8325, COL, MW2, MSSA476, and MRSA252. The number of strains that match either N315 or Mu50 is presented.

[e] Base pairs adjacent to an apparent deletion.

[f] Strain COL possesses the repeat configuration of 38 bp–(2 bp–38 bp) × 9. MW2 and MRSA252 carry 38 bp–(2 bp–38 bp) × 6. In NCTC8325 and MSSA476, no corresponding sequence was found.

[g] The other five strains share the repeat configuration of 9 bp–12 bp–9 bp.

[h] Strain COL possesses the repeat configuration of 199 bp–(185 bp–199 bp) × 4. NCTC8325 carry 199 bp–(185 bp–199 bp) × 7. In both MW2 and MSSA476, the corresponding sequence is missing, yet the flanking sequences form distinct repeats the configuration of which is 227 bp–(157 bp–227 bp) × 6 instead of the configuration of 199 bp–(185 bp–199 bp) × *n*. In MRSA252, neither the corresponding sequence nor any other repeat configuration was found.

together with HsdS of archetypal type IC family members: EcoprrI, EcoR124II, EcoDXXI, Lla1403I, Lla103I, Lla7I, HpyCR2P, HpyCR38P, HpyCR29P, MpnORF342P, NgoAV, and NmeAORF1038P (Adamczyk-Poplawska et al. 2003). Among these, S.Lla1403I, S.Lla103I, and S.Lla7I were selected because they showed high sequence similarity to those HsdSs on genomic islands of *S. aureus* in their conserved regions. Then those sequences were once again aligned by ClustalW and followed by manual refinement to examine their domain structures (see fig. 5*A* below).

The overall structure of *rrn* operon was set by arranging 16S-23S-5S rDNAs, and then the positions of these rDNAs were mapped on the circular genomes (see fig. 6*A* below). The locus names were assigned as illustrated in figure 6*A* (see below). For alleles of the 16S-23S rDNA intergenic spacer region in *rrn* operon, sequences of this region from the seven strains were aligned together with sets of reported sequences of other *S. aureus* strains, H11, ATCC33925, D46, and SAU39769 (Gurtler and Barrie 1995; Forsman, Tilsala-Timisjarvi, and Alatossava 1997). Sequence blocks illustrated in figure 6*C* (see below) were set by ClustalW with manual refinement. The naming for these sequence blocks was renewed here from the former one (Gurtler 1999) to reduce preexisting redundancy.

Results and Discussion

Macroscopic Polymorphisms Between N315 and Mu50 Strains

We identified, in total, 27 macroscopic differences (more than 10 bp) between the genome sequences of two strains, N315 and Mu50. In *S. aureus*, indels of mobile genetic elements and their relatives had been identified, and in many cases, flanking direct repeats were identified at each end of the apparent insert (Kuroda et al. 2001; Baba et al. 2002; Holden et al. 2004; Gill et al. 2005). At first, we confirmed eight indels of entire mobile genetic elements, which include three copies of Tn554, a conjugative transposon Tn5801, two copies of IS1181, a prophage ϕSa1mu, and a genomic island νSa3, which were reported earlier (Kuroda et al. 2001). The mechanisms leading to these polymorphisms are likely to be integration/excision of these mobile genetic elements via transposase-mediated recombination or integrase-mediated site-specific recombination, so they were placed outside the focus of this study.

The remaining 19 sites were examined further. Through examination by CGAT (see *Materials and Methods*), we categorized those polymorphisms into two: six of larger polymorphisms (table 1) and 13 of smaller polymorphisms (table 2). Among the larger polymorphisms, for an indel within serine-aspartate dipeptides repeat region of *clfA* (ID #7) and for many localized polymorphisms in homologous prophages ϕSa3 (ID #13), mechanisms for rearrangements were discussed in the literature (McDevitt and Foster 1995; Foster and Hook 1998; Brussow, Canchaya, and Hardt 2004). Below (*Polymorphisms in Three Tandem Paralogue Clusters Within Genomic Islands, νSaα and νSaβ* and the following sections), we consider detailed mechanisms for the following two, an indel within *ssl* tandem gene cluster in genomic island νSaα (ID #2) and a substitution within *rrn* operon (ID #4), referring to the five genomes other than N315 and Mu50 as well.

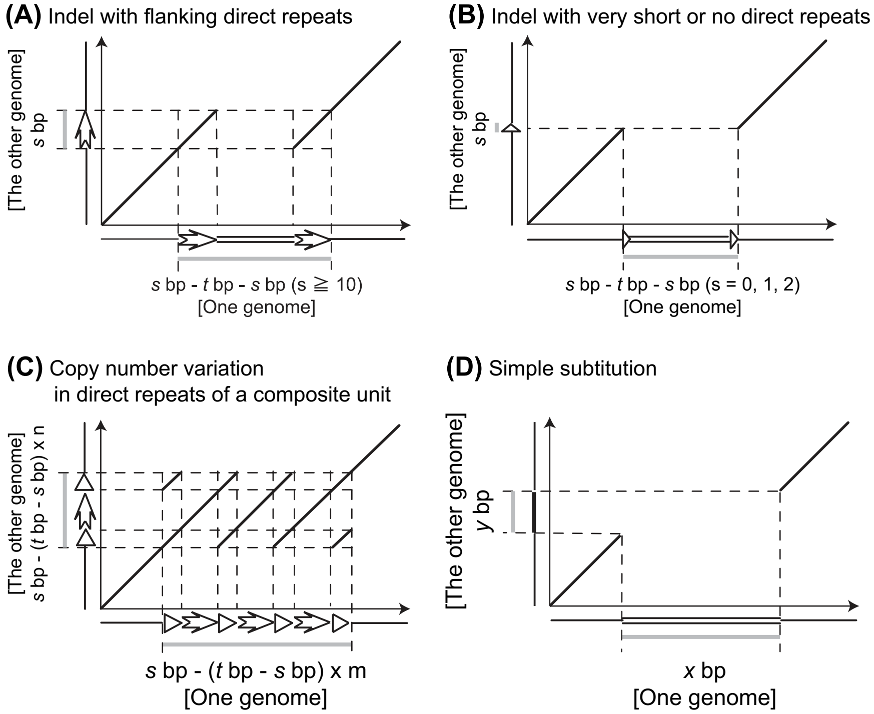


FIG. 1.—Types of smaller genome polymorphisms as classified by dot plot patterns. (*A*) Indel with flanking direct repeats. A part of one genome is apparently duplicated in direct orientation in the other genome, and these repeats flank a DNA segment not present in the former genome. The length of the flanking repeats is equal to or more than 10 bp. (*B*) Indel with very short or no direct repeats. A structure similar to (*A*) except that the length of the flanking repeats is less than 3 bp. (*C*) Copy number variation in direct repeats of a composite unit. A short stretch of DNA sequence is repeated in tandem. The unit of repeats apparently consists of two parts. (*D*) Simple substitution. A part of one genome is replaced by a nonhomologous DNA segment in the other genome. A gray line indicates a polymorphic region.

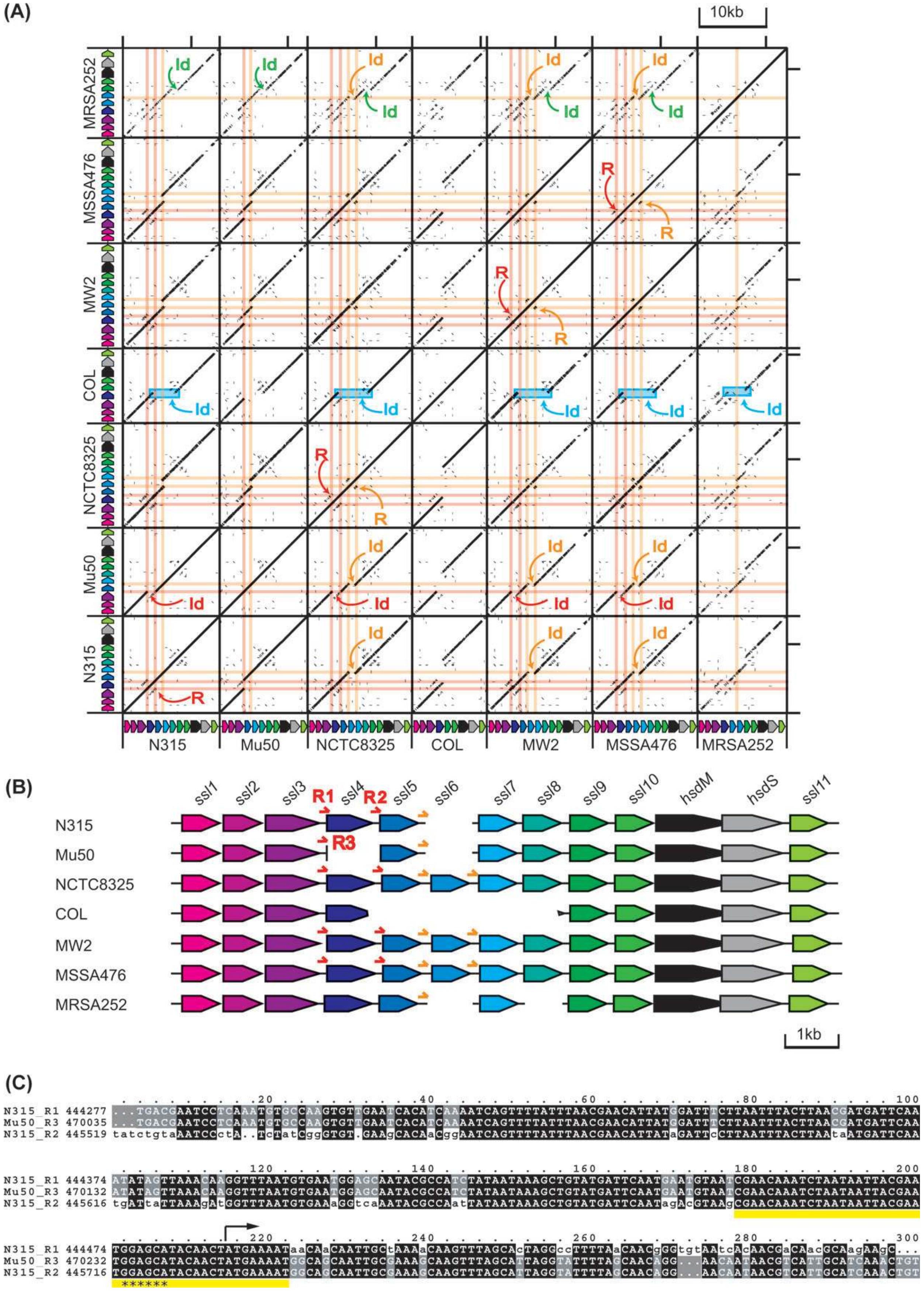
Smaller Polymorphisms Between N315 and Mu50 Strains

These smaller polymorphisms were considered as generated through local events because sequences homologous to these regions were not found elsewhere in the genome in the seven sequenced *S. aureus* strains as analyzed in CGAT (*Materials and Methods*). Based on a dot plot matrix for the local sequences around them (*Materials and Methods*), the 13 smaller polymorphisms were classified into four types (fig. 1 and table 2). Type (A) (fig. 1*A*) consists of indels with simple configuration with flanking direct repeats; a single pair of dispersed repeats in one strain (horizontal axis) has apparently deleted one of the repeats together with the intervening sequence in the other strain (vertical axis). The lengths of the flanking repeats are equal or more than 10 bp. Type (B) (fig. 1*B*) also represents indels, yet, unlike type (A), the length of the flanking direct repeats, if any, is less than 3 bp. The threshold between types (A) and (B) is set considering their possible mechanisms (see *Criteria for Classification of Putative Rearrangement Events* in *Materials and Methods*). Type (C) (fig. 1*C*) can also be considered as indels with flanking direct repeats. However, in this type, a sequence is repeated multiple times (copy number >2), and the copy number varies between the two genomes. The unit of the repeats apparently consists of two parts. Type (D) (fig. 1*D*), simple substitution, is seen in the polymorphism ID #8.

What are likely mechanisms leading to these polymorphisms? The indels of types (A) and (C) can be explained by recombination events between the flanking direct repeats. In order to distinguish between homologous recombination and illegitimate recombination, we examined the lengths and sequence identities of the flanking direct repeats (table 2). For the polymorphism ID #3 and ID #19 with 188-bp homology and 384-bp homology, respectively, homologous recombination appears likely, although *recA*-independent replication error is not excluded. The remainder is likely to have been generated by various types of illegitimate recombination requiring sequence similarity. For type (B), the lengths of the flanking repeats, if any, are insufficient for neither homologous recombination nor illegitimate recombination involving short repeats. Illegitimate recombination between sequences with no or very little similarity can explain these polymorphisms.

Table 2 also refers to the corresponding regions of these polymorphic loci in the other five strains than N315 and Mu50. Notably, in all the cases of types (A) and (B), all the other five strains matched either N315 or Mu50, that is, the N315 or Mu50 showed an exceptional pattern among the seven sequenced strains. Furthermore, the exception, whether it be N315 or Mu50, is the shorter type in all the cases. This directionality indicates that these polymorphisms were generated by deletion, as opposed to insertion, in either N315 or Mu50 strain lineage.

On the other hand, with type (C), there is no such simple tendency as seen in the types (A) and (B). Instead, in these polymorphisms, several of the other five strains possess the same repeat unit although they differ in the repeat


(A)
10kb
MRSA252
MSSA476
MW2
COL
NCTC8325
Mu50
N315
Id
R
N315
Mu50
NCTC8325
COL
MW2
MSSA476
MRSA252
(B)
ssl1
ssl2
ssl3
R1
ssl4
R2
ssl5
ssl6
ssl7
ssl8
ssl9
ssl10
hsdM
hsdS
ssl11
R3
N315
Mu50
NCTC8325
COL
MW2
MSSA476
MRSA252
1kb
(C)
. . . . .20 . . . . .40 . . . . .60 . . . . .80 100
N315_R1 444277 ...TGACGAATCCTCAAATGTGCCAAGTGTTGAATCACATCAAAATCAGTTTATTTAACGAACATTATGGATTTCTTAATTTACTTAACGATGATTCAA
Mu50_R3 470035 ...TGACGAATCCTCAAATGTGCCAAGTGTTGAATCACATCAAAATCAGTTTATTTAACGAACATTATGGATTTCTTAATTTACTTAACGATGATTCAA
N315_R2 445519 tatctgtaAATCCctA..TcTatCggGTGT.GAAgCACAaCggAATCAGTTTATTTAACGAACATTATaGATTcCTTAATTTACTTAAtaATGATTCAA
. . . . 120 . . . . 140 . . . . 160 . . . . 180 200
N315_R1 444374 ATATAGTTAAACAAGGTTTAATGTGAATGGAGCAATACGCCATCTATAATAAAGCTGTATGATTCAATGAATGTAATCGAACAAATCTAATAATTACGAA
Mu50_R3 470132 ATATAGTTAAACAAGGTTTAATGTGAATGGAGCAATACGCCATCTATAATAAAGCTGTATGATTCAATGAATGTAATCGAACAAATCTAATAATTACGAA
N315_R2 445616 tgATtaTTAAAgAtGGTTTAATGTGAAaGGtcaAATACGCCAatTATAATAAAGCTGTATGATTCAATagAcGTAAgCGAACAAATCTAATAATTACGAA
. . . . 220 . . . . 240 . . . . 260 . . . . 280 300
N315_R1 444474 TGGAGCATACAACTATGAAAATaaCAaCAATTGCtAAAaCAAGTTTAGCAcTAGGccTTTTAaCAACgGGtgtAAtcAcAACGaCAacGCAagAAgC...
Mu50_R3 470232 TGGAGCATACAACTATGAAAATGGCAGCAATTGCGAAAGCAAGTTTAGCATTAGGTATTTTAGCAACAGG...AACAATAACGTCATTGCATCAAACTGT
N315_R2 445716 TGGAGCATACAACTATGAAAATGGCAGCAATTGCGAAAGCAAGTTTAGCATTAGGTATTTTAGCAACAGG...AACAATAACGTCATTGCATCAAACTGT
******

number (see footnotes of table 2). These observations can be interpreted in terms of generation of a composite unit of the type *s-t-s* (see fig. 1*C*) and its expansion and contraction through recombination involving the repeats.

Polymorphisms in Three Tandem Paralogue Clusters Within Genomic Islands, νSaα and νSaβ

ssl Cluster

Our initial pairwise comparison between N315 and Mu50 genomes confirmed an indel in *ssl* gene cluster (table 1, ID #2), which results in loss of one ORF, *ssl4* in figure 2*B*, in Mu50 (Kuroda et al. 2001; Fitzgerald et al. 2003) (see also *Introduction*).

Figure 2*A* represents multiple dot plots with these regions from the seven sequenced genomes. The successive lines with some breaks in the diagonal part of each rectangle for a pairwise comparison indicate conserved gene order with occasional indels. The presence of only few black lines or dots other than the diagonal one in each rectangle for self-to-self plot is consistent with the divergence of the tandem genes within a strain. Four groups of indels were identified in all in the dot plots for pairwise comparison (*Materials and Methods*) among seven genomes and are labeled Id in red, orange, blue, and green, respectively, in figure 2*A*.

Through each of the two break points of the red indel, a horizontal line and a vertical line are drawn across the entire plots in red in figure 2*A*. In the self-to-self plots there, dots are visible at the cross section of two red lines. These indicate direct repeats flanking the red indel. The direct repeats flanking the orange indel were identified similarly with the help of orange lines (fig. 2*A*). These indels likely represent deletions that have been generated by recombination between these flanking direct repeats.

The positions of these repeats are shown on the maps of this gene cluster in figure 2*B* as red arrows and as orange arrows, which reveals that the relative position of the repeat to the closest ORF is nearly identical for the two repeats. The map also shows conservation of the repeats among the strains.

Through a multiple alignment of the relevant recombination joints (with ClustalW, *Materials and Methods*), we defined repeats flanking each indel in terms of length and sequence identity. Figure 2*C* shows the repeats flanking the red indel (fig. 2*A* and *B*) between N315 and Mu50. The 45-bp perfect match is seen at the putative recombination point (underlined in yellow in fig. 2*C*), and the length of homology could be elongated to 108 bp (444388–444495 [N315], 445630–445737 [N315], and 470146–470253 [Mu50]) if 10% sequence divergence was allowed. This 45-bp sequence codes for the initiating ATG, the putative ribosome-binding site, and a part of the putative signal peptides for secretion (Williams et al. 2000).

For the orange indel in figure 2*A* and *B*, the length of the perfect match was 61 bp (5′-GTGACATGAAACAATGTGGAAAACATAATTAAATTGAGGGAAAGTGTGAATAGTTAAAAAA-3′) between N315 and MW2, and the repeats' length could be elongated to 84 bp if 10% divergence was allowed (434809–434892 [MW2], 435929–436012 [MW2], and 446612–446695 [N315]). There was a 184-bp sequence between this repeated sequence and the nearest ORF to its right in the case of N315.

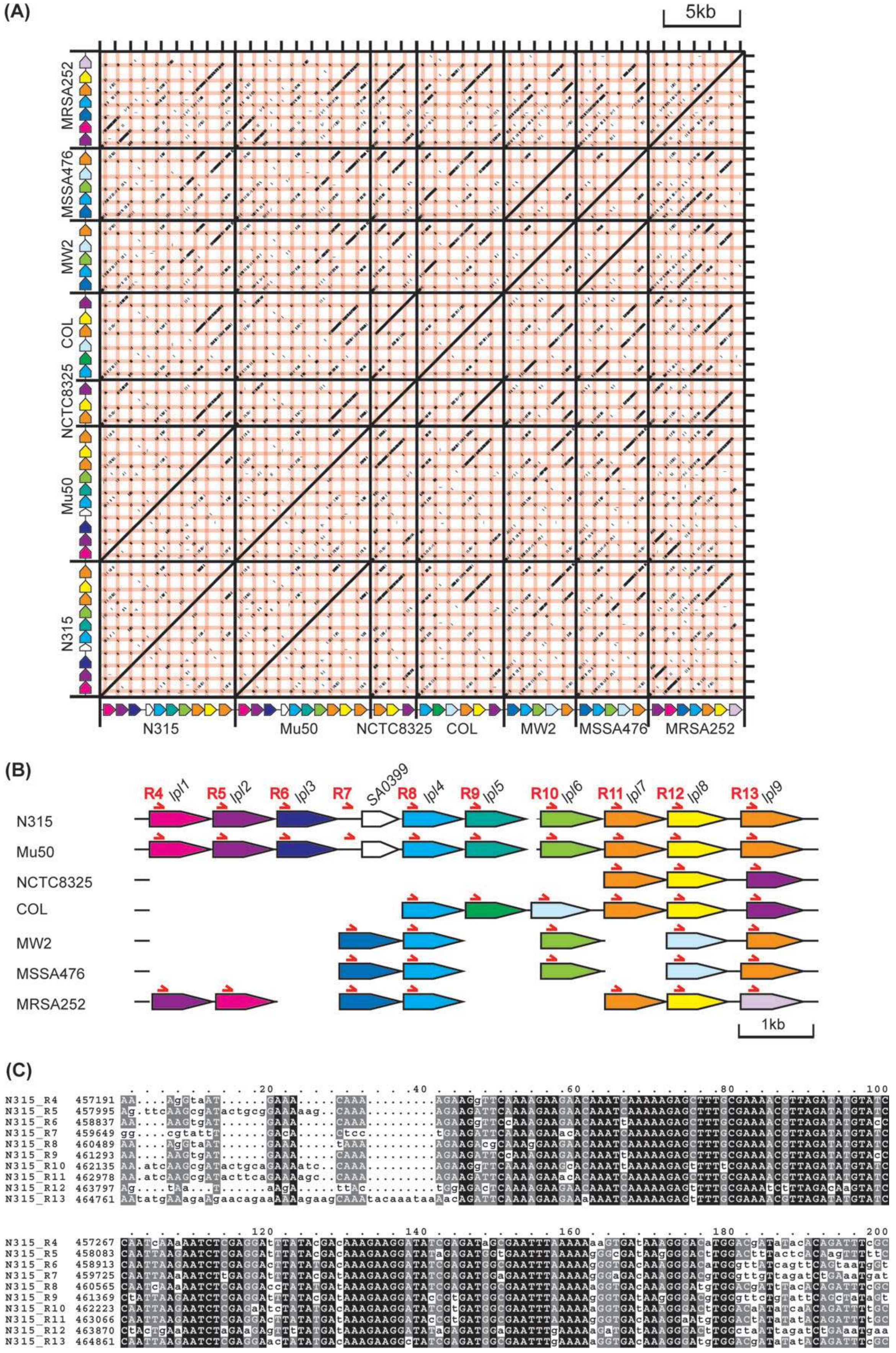
A similar analysis of the remaining two indels (in blue and green in fig. 2*A*) revealed only very short flanking direct repeats, which are insufficient for homologous recombination. Even if 10% sequence divergence was allowed, only 16-bp-long repeats were detected between NCTC8325 and COL for the apparent deletion (in blue) in COL (392547–392564 [NCTC8325], 396915–396931 [NCTC8325], and 474767–474783 [COL]). Only 13-bp-long repeats were detected between MSSA476 and MRSA252 for the apparent deletion (in green) in MRSA252 (435859–435871 [MSSA476], 436936–436948 [MSSA476], and 459325–459337 [MRSA252]).

lpl Cluster

Similar tandem gene clusters, *lpl* cluster on νSaα and *spl* cluster on νSaβ (see *Introduction*) were examined for mechanisms responsible for interstrain difference. Contrary to the *ssl* cluster, the same multiple dot plot analysis of the *lpl* cluster (fig. 3*A*) yielded multiple discrete lines in parallel to a diagonal line in most of the rectangles for pairwise comparison, which suggests extensive genome rearrangements. We noticed that many of their endpoints mapped at a specific site in all the ORFs. This became apparent when a vertical line and a horizontal line were drawn through this site for all the ORFs. This feature indicates that the highly identical repeats exist at the boundaries of the homology and nonhomology. Furthermore, many of the cross sections between these red lines coincided with a (black) dot in figure 3*A*. This lattice pattern of the dots implies that this sequence is repeated within a genome as well as between genomes.

←

FIG. 2.—Comparison of the staphylococcal superantigen-like (*ssl*) gene clusters of seven sequenced *Staphylococcous aureus* strains. (*A*) Dot plots. The sequences from *ssl1* through *ssl11* with additional 200 bp at each end were plotted against one another. The breaks that specify four examples of indels identified here are labeled Id in green, orange, blue, and red, respectively. The large, blue indel is highlighted by a blue rectangle. The indels shown in red and orange are indels with flanking direct repeats. The dots that specify these flanking intragenomic repeats are labeled R in respective colors. Through each of the two break points of the red indel, a horizontal line and a vertical line are drawn across the entire plot in red. The orange lines are drawn in the same way. (*B*) ORF map. The above repeats are displayed in more detail as red and orange arrows. The naming and alignment of *ssl* genes are after International Nomenclature Committee for Staphylococcal Superantigens (Fitzgerald et al. 2003; Lina et al. 2004), respectively. Note that a boundary of homology and nonhomology at DNA level is located within an ORF for the left deletion in Mu50 and the deletion in COL. (*C*) Alignment of the direct repeats R1, R2, and R3 in (*B*) suggesting a recombination event between them. In the upstream region, R3 perfectly aligns with R1 (as indicated by the gray shading), but not with R2, then shares 45-bp sequence (underlined in yellow) with both R1 and R2, and then perfectly aligns with R2 (as indicated by the gray shading), but not with R1. The initiating ATG is indicated by an upper arrow, and the putative ribosome-binding site (Williams et al. 2000) is indicated by asterisks.



```
                  . . . . . . . . .20 . . . . . . . . .40 . . . . . . . . .60 . . . . . . . . .80 . . . . . . . . 100
N315_R4  457191 AA....AgGtaAT......GAAA.....CAAA.........AGAAGgTTCAAAAGAAGAACAAATCAAAAAGAGCTTTGCGAAAACGTTAGATATGTATC
N315_R5  457995 Ag.ttcAAGcgATactgcgGAAAaag..CAAA.........AGAAGATTCAAAAGAAGAACAAATCAAAAAGAGCTTTGCGAAAACGTTAGATATGTATC
N315_R6  458837 AA....AAGtgAT......GAAA.....CAAA.........AGAAGgTTCcAAAGAAGAACAAATtAAAAAGAGCTTTGCGAAAACGTTAGATATGTACC
N315_R7  459649 gg....cgtattT......GAcA.....Ctcc.........tGAAGATTCAAAAGAAacACAAATCAAAAAGAGCTTTGCGAAAACGTTAGATATGTATC
N315_R8  460489 AA....AgGtaAT......GAAA.....tAAA.........AGAAGAcgCAAAgGAAGAACAAATCAAAAAGAGCTTTGCGAAAACGTTAGATATGTATC
N315_R9  461293 AA....AAGtgAT......GAAA.....CAAA.........AGAAGATTCcAAAGAAGAACAAATtAAAAAGAGCTTTGCGAAAACGTTAGATATGTACC
N315_R10 462135 AA.atcAAGcgATactgcaGAAAatc..CAAA.........AGAAGgTTCAAAAGAAGcACAAATtAAAAAGAGtTTTtCGAAAACGTTAGATATGTATC
N315_R11 462978 AA.atcAAGcgATacttcaGAAAagc..CAAA.........AGAAGATTCAAAAGAAacACAAATCAAAAAGAGCTTTGCGAAAACGTTAGATATGTATC
N315_R12 463797 Ag....tAa...T......aAgA.....ttAc.........tGgAGAcTCAAAAGAAGAACAAATCAAAAAGAGtTTTGCGAAAtCtTTAGAcAaGTATC
N315_R13 464761 AAtatgAAagaAgaacagaaAAAagaagCAAAtacaaataaAacAGATTCAAAAGAAGAAaAAATCAAAAAGAGtTTTGCGAAAACGTTAGATATGTATC

                  . . . . . . . . 120 . . . . . . . . 140 . . . . . . . . 160 . . . . . . . . 180 . . . . . . . . 200
N315_R4  457267 CAATcAAaAATCTCGAGGAtTTATAcGAcAAAGAAGGATATCGAGATaGCGAATTTAAAAAaaGTGAtAAAGGGACaTGGACgATaTacACAGATTTcGC
N315_R5  458083 CAATTAAGAATCTCGAGGAtTTATAcGAcAAAGAAGGATATaGAGATGGtGAATTTAAAAAgGGcGAtAAgGGGACtTGGACttTactcACAagTTTttC
N315_R6  458913 CAATTAAGAATCTCGAGGAcTTATATGAcAAAGAAGGATATCGAGATGGCGAATTTAAAAAgGGTGAcAAAGGGACaTGGgttATcagttCAGtaaTgGt
N315_R7  459725 CAATTAAaAATCTtGAGGAtTTATAcGAtAAAGAAGGATATCGAGATGGaGAATTTAAAAAgGGaGAcAAAGGGACgTGGgttgTtagatCtGAaaTgat
N315_R8  460565 CAATcAAaAATCTCGAGGAccTATATGAtAAAGAAGGATATCGAGATGGCGAATTTAAAAAaGGTGAcAAAGGGAtgTGGACgATtTacACAGATTTcGC
N315_R9  461369 CtATTAAGAATCTCGAGGAtTTATAcGAtAAAGAAGGATAcCGtGATGGCGAATTTAAAAAgGGTGAtAAgGGGACgTGGgttcTgTattCAGcTaTaGt
N315_R10 462223 CAATTAAGAATCTCGAGaAtcTATATGAcAAAGAAGGATATCGtGATGGCGAATTTAAAAAaGGTGAtAAAGGGACtTGGACaATaTcaACAGATTTtGC
N315_R11 463066 CAATTAAGAATCTCGAGGAcTTATATGAtAAAGAAGGATAcCGtGATGGCGAATTTAAAAAgGGTGAcAAAGGaAtgTGGACtATaTatACAGATTTtGC
N315_R12 463870 CtAcTgAaAATCTaGAaGAgTTtTATGAtAAAGAAGGATATaGAGATGGaGAATTTgAAAAaGaTGAtAAAGGGACtTGGctaATtagatCtGAaaTgaa
N315_R13 464861 CAATTAAGAATCTCGAGGAacTATATGAcAAAGAAGGcTATCGAGATGGCGAATTTgAAAAaGGTGAcAAAGGGAtgTGGACgATaTatACAGATTTcGC
```

These *lpl* ORFs were classified into 12 phylogenetic groups (*Materials and Methods*) displayed in distinct colors in figure 3*A* and *B*. Their maps in figure 3*B* revealed variation in the gene order, which suggests extensive genome rearrangements. Individual rearrangements such as indels, duplications, or translocations were hard to identify in the pairwise comparison.

Figure 3*B* also shows the above repeats on ORF maps. All the copies of this repeated sequence are located at a similar position near the 5′ terminus of ORF. Notably, in N315 and Mu50, the dots indicating these patterned repeats are also seen upstream of the ORFs (colored in white; SA0399 for N315) not annotated as a lipoprotein (displayed as R7 for N315 in fig. 3*B*). These two ORFs may represent remnants of the 3′ part of an *lpl* gene. In support of this view, the deduced amino acid sequence for these ORFs aligned well with the C-terminal region of other *lpl* genes (data not shown).

The repeated sequences in N315 are aligned in figure 3*C*. In the 27 pairs among all the possible pairs (45), the repeated length is 80 bp or more if 10% divergence is allowed. Judging from the lengths and sequence identities between them, homologous recombination likely has occurred between these repeats to generate the observed gene order changes.

spl cluster

Similar dot plots for the *spl* cluster are shown in figure 4*A* together with ORF maps in figure 4*B*. The *spl* ORFs were classified into seven phylogenetic groups (*Materials and Methods*) displayed in distinct colors in figure 4*A* and *B*. The right part of this cluster seems to be conserved between all the examined strains except for MRSA252, in which the yellow ORF (*splB* in N315) is truncated, presumably by another rearrangement event, leaving its 3′ terminus (SAR1907). In the left half, some rearrangements are seen among the seven strains. Their simplest description would be as follows. An ORF in violet (*splF* in N315) in MW2, MSSA476, and MRSA252 appears duplicated in N315, Mu50, NCTC8325, and COL. An ORF in blue is present in NCTC8325, COL, and MRSA252 but not in the other four strains.

The dot plot patterns corresponding to these two types of polymorphisms are labeled as Dp and Id, respectively, in figure 4*A*. As in the *lpl* cluster, many of the endpoints of the discrete lines corresponding to these two groups of rearrangements mapped at 5′ terminus of several ORFs (fig. 4*A*). This became apparent when a vertical line and a horizontal line were drawn through these sites. Furthermore, some of the cross sections between these red lines coincided with a (black) dot in figure 4*A*. This pattern implies that this sequence is repeated within a genome. These results suggest that these repeats are associated with those rearrangements via some homology-dependent recombination events. These very similar repeats present at the recombination joints are drawn as arrows in figure 4*B*.

In search for the sequences involved in the recombination event, a multiple alignment was computed for the indel seen between N315 and COL as an example (fig. 4*C*). The length of the perfect match is 12 bp (underlined in blue in fig. 4*C*), and the repeats length could be elongated to 223 bp covering the initiating ATG and the 108 nucleotides encoding the signal peptides (Reed et al. 2001) (underlined in gray in fig. 4*C*) if 10% divergence was allowed (1918257–1918036 [COL], 1917391–1917169 [COL], and 1861882–1861661 [N315]). It was therefore suggested that this indel represents a deletion caused by homologous recombination between the repeats.

Polymorphisms in Type I RM System Genes Within Genomic Islands, νSaα and νSaβ

An *hsdM* and *hsdS* gene pair is found in both νSaα and νSaβ genomic islands of the seven strains. A similar *hsdM* and *hsdS* gene pair was identified in another genomic island, *etd*, of *S. aureus* strain TY114 (Yamaguchi et al. 2002). Earlier work pointed out that the deduced amino acid sequences for *hsdM* are almost identical among all the 15 alleles, but those for *hsdS* show divergence into seven different types with most of the differences localized in the target sequence recognition domains (Baba et al. 2002). We carried out more detailed sequence comparison to characterize the variation of these *hsdS* genes.

Type I RM genes have been grouped into five families, type IA through type IE (Murray 2000; Chin et al. 2004), and some of the above genes were already grouped into type IC family (Kuroda et al. 2001). To ascertain their domain structure and to characterize their variation, we compared their sequences with those of other type I RM genes (see *Materials and Methods*). Earlier works revealed that all the members of one family show strong homology in the conserved regions of HsdS (Murray 2000), but recent sequence data have revealed that those similarities are lower than estimated for the prototype members (Titheradge et al. 2001; Adamczyk-Poplawska et al. 2003).

Figure 5*A* shows a multiple alignment of the seven types of HsdS proteins of *S. aureus* and other three types of HsdS proteins of type IC members of *Lactococcus lactis*. After manual refinement, three conserved regions (designated as N-terminal, central, and C-terminal conserved regions) and two intervening variable regions that are likely responsible for target sequence recognition (designated as N-terminal and C-terminal target recognition domains)

←

FIG. 3.—Comparison of the putative lipoprotein (*lpl*) gene clusters of seven sequenced *Staphylococcus aureus* strains. (*A*) Dot plots. The sequences from *lpl1* through *lpl9* in strain N315 with additional 200 bp at each end and the corresponding genomic regions in the other strains were plotted against one another. The horizontal and vertical lines in red indicate a specific conserved sequence in each ORF that served as the recombination site for many rearrangements. (*B*) ORF map. The above sequence is shown as a red arrow. Color was assigned to each ORF based on phylogenetic grouping. The gene names are those for N315 (Kuroda et al. 2001). A white arrow represents a hypothetical ORF. (*C*) Alignment of R4 through R13 of strain N315 illustrated in (*B*).

were identified, as was reported for various HsdSs (Titheradge et al. 2001). The sequences of the conserved regions aligned very well within the HsdSs of *S. aureus* and within those of *L. lactis*, respectively. There is also some interspecific similarity (fig. 5*A*).

Figure 5*B* shows schematic diagrams for the organization of the seven types of HsdS proteins of *S. aureus*. Reassortment of the two target recognition domains can be identified in intrastrain and interstrain comparison: type 1 and type 3 share the sequence of the C-terminal target recognition domain, while they do not share that of the N-terminal target recognition domain. Likewise type 2, type 4, and type 5 share the sequence of the N-terminal target recognition domain but not that of the C-terminal one.

The reassortment of the two target recognition domains was shown to create novel target specificity (Fuller-Pace et al. 1984; Gann et al. 1987; Gubler et al. 1992). The present result suggests that such reassortment occurred also under natural conditions. Furthermore, it is inferred that this reassortment was caused by some mode of homologous recombination involving the central conserved region. Indeed, the nucleotide sequence alignment of the central conserved region of the seven types of *hsdS* in *S. aureus* showed that there is a common sequence as long as 140 bp sharing more than 90% in 18 out of all (21) the possible combinations (data not shown).

Polymorphisms in *rrn* Operons

Figure 6*A* and *B* shows copy number, chromosomal position, gene organization, and orientation of ribosomal RNA operons in the seven genomes. There are two types of strains, the strains with five *rrn* copies (N315, Mu50, NCTC8325, and MRSA252) and those with six *rrn* copies (COL, MW2, and MSSA476), as reported (Kuroda et al. 2001; Baba et al. 2002; Holden et al. 2004; Gill et al. 2005). The gene organization is 16S-23S-5S in all the cases except for an additional 5S rDNA upstream of locus 2 and locus 2-1. In the strains with six copies, two copies of *rrn* operons on locus 2-1 and 2-2 are found located next to each other intervened only by a 211-bp sequence (between 5S rDNA of locus 2-1 and 16S rDNA of locus 2-2). This corresponds to the single copy on locus 2 of the strains with five copies, although the 107-bp sequence (black box in fig. 5*B*) between the 3′ end of the 5S rDNA of locus 2-1 and the point 104-bp upstream of the 16S rDNA of locus 2-2 was not found in locus 2.

Our initial comparison between N315 and Mu50 revealed one substitution in a 16S-23S rDNA intergenic spacer region (table 2, ID #3). Figure 6*C* illustrates the sequence patterns of 16S-23S intergenic spacer region found in the seven strains, which show a mosaic structure that consists of three conserved sequence blocks, CS1 through CS3, and different sets of variable sequence blocks, VS1 through VS13 (see legend for fig. 6*C*). The upper nine alleles (*rrnA-rrnH*) were already reported (Gurtler and Barrie 1995; Forsman, Tilsala-Timisjarvi, and Alatossava 1997). The last two were discovered in this study from MW2 and MSSA476 and MRSA252 and designated as *rrnY* and *rrnZ*, respectively. While *rrnY* consists of the known variable sequence blocks in a novel combination, *rrnZ* bears a novel 16-bp sequence, 5′-GTGATAA-TAAAGCAGT-3′, designated as VS13, apparently inserted within VS4. This 16-bp sequence was also found at the locus 4 in MRSA252, which was caused by base substitution mutations from the consensus sequence of *rrnH*, 5′-aTGAaAAATAAAGCAGT-3′, comprising 3 bp of VS4 and 13 bp of VS5.

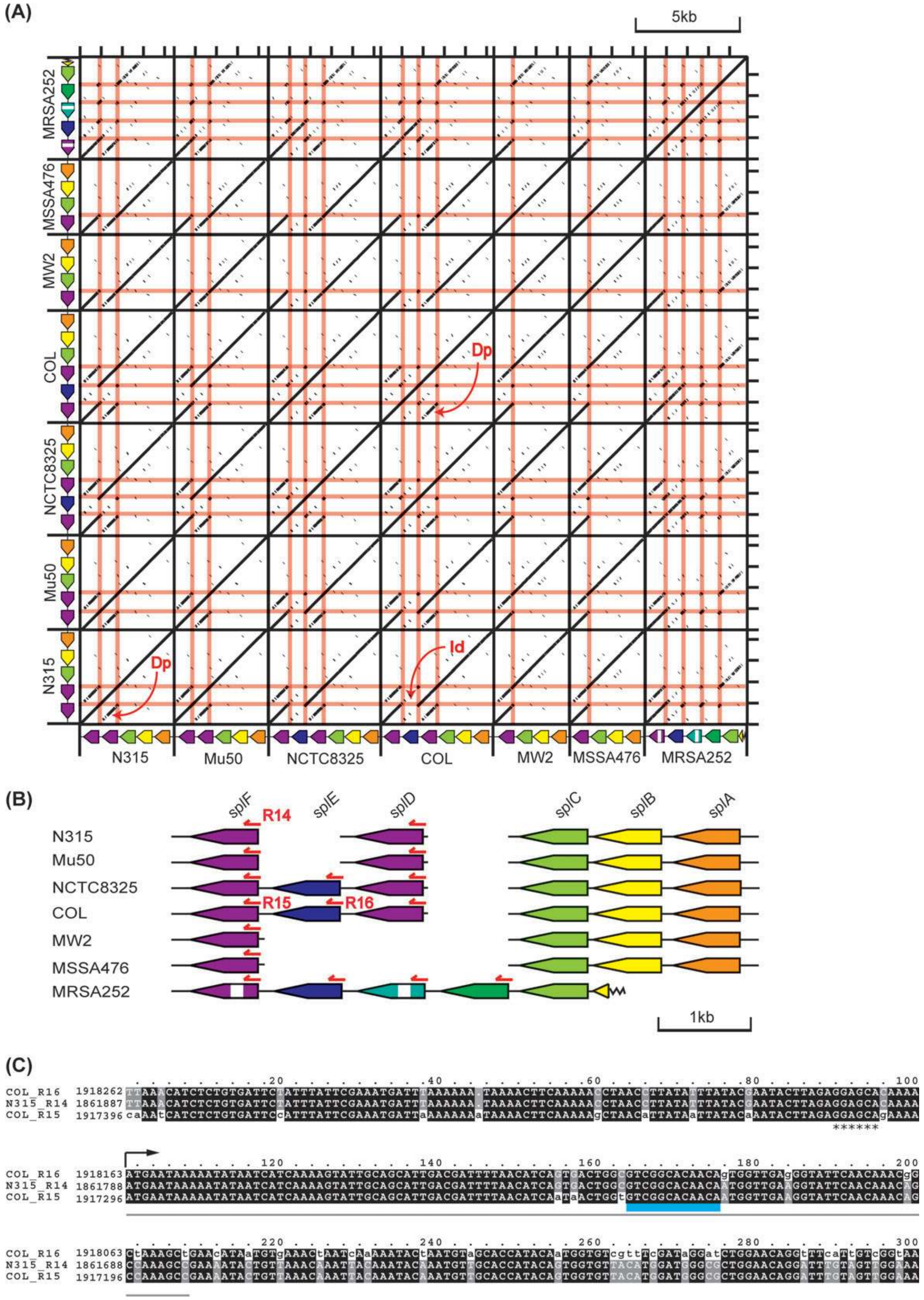
Table 3 shows which of the above types of *rrn* operon is present in each locus of each strain. Between N315 and Mu50, the only difference is on the locus 1, which is apparently a simple substitution of VS1-VS2-VS3 by VS4-VS6-VS7. However, *rrnC* is found both on the locus 1 and the locus 3 in N315. This could be a result of gene conversion in the N315 lineage involving *rrnH* at locus 1 in the Mu50-type ancestor as the recipient and *rrnC* at locus 3 as the donor. A similar situation is seen between NCTC8325 and COL for locus 4 and locus 5, from *rrnF* to A48073. Another interesting relationship is found between MW2 and MSSA476; the locus 2-1 and the locus 2-2 are apparently reciprocally exchanged between *rrnK* and *rrnJ*.

What kind of mechanisms can be inferred for them? The indel involving whole *rrn* operons observed at loci 2-1 and 2-2 and locus 2 can be explained by an unequal crossing-over leading to deletion or tandem duplication. The former deletion can occur by a single homologous recombination event via extensive homology of duplication. For the latter duplication process, unequal crossing-over between the flanking 5S rDNA genes at locus 2 is not sufficient. Such a homologous recombination event should be followed by another illegitimate recombination event in order to explain the presence of the 107-bp sequence between locus 2-1 and locus 2-2, which is absent from locus 2.

The various *rrn* alleles in the 16S-23S intergenic spacer region can be explained as combinations of a left variable region and a right variable region (fig. 6*C*). This recombination is likely through homologous recombination involving the long homology of CS1 and CS2.

→

FIG. 4.—Comparison of the putative serine protease (*spl*) gene clusters of seven sequenced *Staphylococcus aureus* strains. (*A*) Dot plots. The sequences from *splA* through *splF* in strain N315 and the homologous genomic regions in the other strains were plotted against one another. Id indicates one of the patterns representing indel of *splE* homologue in blue. Dp indicates one of the patterns representing duplication of *splF*/*splD* homologue in violet. Horizontal and vertical lines in red are drawn through their break points. (*B*) ORF map. The sequence corresponding to the break points of the above rearrangements is shown as a red arrow. Color was assigned to each ORF based on phylogenetic grouping. The gene names are those for N315 (Kuroda et al. 2001). An arrow with a white box represents a frameshifted ORF, and a yellow triangle represents a truncated ORF. Nonhomologous 200-bp region in the right end of MRSA252 is indicated by a zigzag line. (*C*) Alignment of R14, R15, and R16 in (*B*) indicating a recombination event between the intragenomic repeats. R14, which perfectly aligns with R16, but not with R15, in the upstream region (gray shading) comes to be so with R15, but not with R14, in the downstream region (gray shading). The blue underline indicates 12-bp perfect identity shared at the transition region. The putative ribosome-binding site is indicated by asterisks. The initiating ATG is indicated by an upper arrow. The 108 nucleotides encoding the signal peptides (Reed et al. 2001) are underlined.



```
                 . . . . . . . . . .20 . . . . . . . . . .40 . . . . . . . . . .60 . . . . . . . . . .80 . . . . . . . . . 100
COL_R16  1918262 TTAAACATCTCTGTGATTCTATTTATTCGAAATGATTTAAAAAA.TAAAACTTCAAAAACCTAACCTTATATTTATACGAATACTTAGAGGAGCACAAAA
N315_R14 1861887 TTAAACATCTCTGTGATTCTATTTATTCGAAATGATTTAAAAAA.TAAAACTTCAAAAACCTAACCTTATATTTATACGAATACTTAGAGGAGCACAAAA
COL_R15  1917396 caAAtCATCTCTGTGATTCcATTTATTCGAAATGATTaAAAAAAaTAAAACTTCAAAAAgCTAACaTTATAaTTATACaAATACTTAGAGGAGCAgAAAA
                                                                                                     ******

                 . . . . . . . . . 120 . . . . . . . . . 140 . . . . . . . . . 160 . . . . . . . . . 180 . . . . . . . . . 200
COL_R16  1918163 ATGAATAAAAATATAATCATCAAAAGTATTGCAGCATTGACGATTTTAACATCAGTGACTGGCGTCGGCACAACAgTGGTTGAgGGTATTCAACAAACgG
N315_R14 1861788 ATGAATAAAAATATAATCATCAAAAGTATTGCAGCATTGACGATTTTAACATCAGTGACTGGCGTCGGCACAACAATGGTTGAAGGTATTCAACAAACAG
COL_R15  1917296 ATGAATAAAAATATAATCATCAAAAGTATTGCAGCATTGACGATTTTAACATCAaTaACTGGtGTCGGCACAACAATGGTTGAAGGTATTCAACAAACAG

                 . . . . . . . . . 220 . . . . . . . . . 240 . . . . . . . . . 260 . . . . . . . . . 280 . . . . . . . . . 300
COL_R16  1918063 CtAAAGCtGAAcATAaTGTgAAACtAATcAaAAATACtAATGTaGCACCATACAaTGGTGTcgttTcGATaGGatCTGGAACAGGtTTcaTtGTcGGtAA
N315_R14 1861688 CCAAAGCCGAAAATACTGTTAAACAAATTACAAATACAAATGTTGCACCATACAGTGGTGTTACATGGATGGGCGCTGGAACAGGATTTGTAGTTGGAAA
COL_R15  1917196 CCAAAGCCGAAAATACTGTTAAACAAATTACAAATACAAATGTTGCACCATACAGTGGTGTTACATGGATGGGCGCTGGAACAGGATTTGTAGTTGGAAA
```

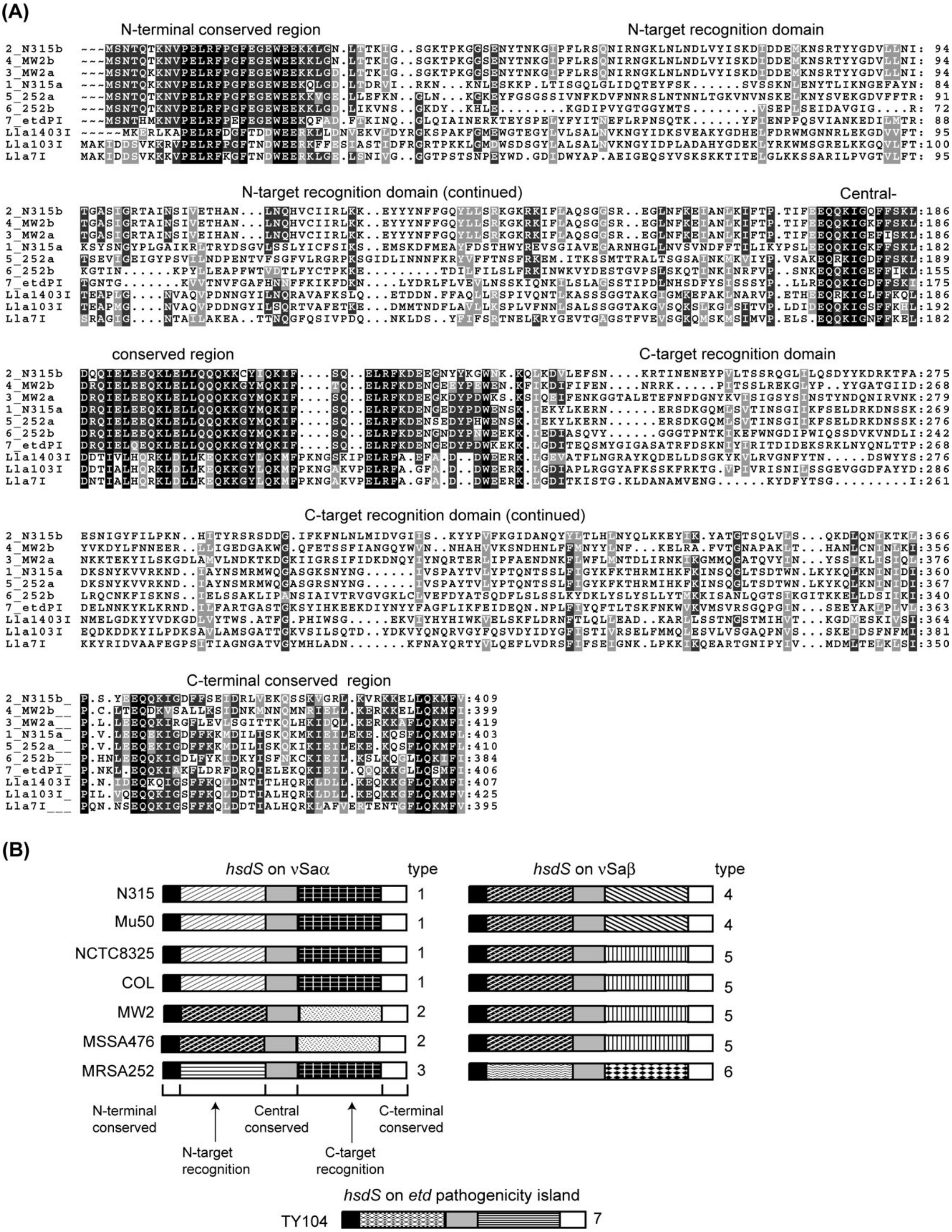


FIG. 5.—Polymorphisms in *hsdS* gene. (*A*) Alignment of the predicted amino acid sequences of HsdS. Representatives of the seven different types 1 through 7 from *Staphylococcus aureus* are compared with each other, together with three type IC HsdS in *Lactococcus lactis* for reference. The sequence name 2_N315b indicates, for example, the sequence from νSaβ of strain N315 as a representative of type 2, and Lla1403I indicates that of S.Lla1403I. (*B*) Schematic diagrams of organization of *hsdS* in seven sequenced *S. aureus* strains and its putative homologue in *etd* pathogenicity island in strain TY104.

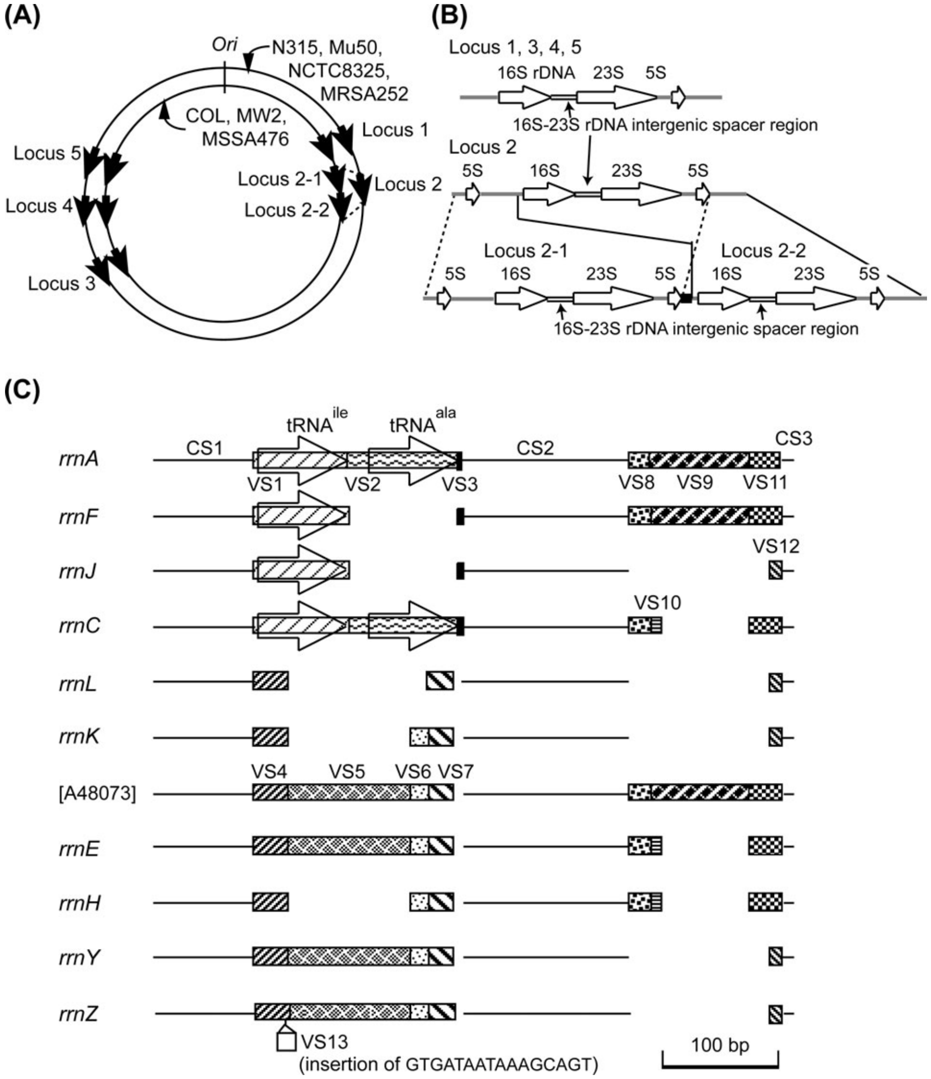


FIG. 6.—The *rrn* operon of seven sequenced *Staphylococcus aureus* strains. (*A*) Location. Each strain has five or six copies of *rrn* operon, the loci of which are numbered clockwise from the predicted replication origin (*Ori*). A black arrow indicates an *rrn* operon in the orientation of 16S-23S-5S. Note that locus 2-1 and locus 2-2 in the strains with six copies are next to each other and correspond to locus 2 in the strains with five copies. (*B*) Organization. Only rRNA-coding regions are displayed. A black box indicates a 107-bp sequence between locus 2-1 and locus 2-2, which is absent from locus 2. (*C*) Alleles of the 16S-23S rDNA intergenic spacer region in *rrn* operon. Sequences of this region from the seven strains were aligned together with sets of sequences that had been reported on other *S. aureus* strains, H11, ATCC33925, D46, and SAU39769 (Gurtler and Barrie 1995; Forsman, Tilsala-Timisjarvi, and Alatossava 1997). The names of the operons are from literature (Gurtler and Barrie 1995) except for A48073, from the GenBank accession number, and the last two, *rrnY* and *rrnZ*, newly added in this study. See table 3 for location of these alleles in the *rrn* loci. The DNA sequences in this region consist of three conserved sequences, CS1 to CS3, and 13 variable sequences, VS1 through VS13. VS1 and VS2 encode tRNA (Ile) and tRNA (Ala), respectively. The consensus lengths (bp) of CS1, CS2, and CS3 are 87, 144, and 10, respectively, and those of VS1 through VS13 are 82, 94, 4, 30, 105, 16, 22, 19, 95, 16, 20, 9, and 16, respectively. The naming of these sequence blocks was modified here from the former one (Gurtler 1999) to reduce preexisting redundancy. Sequence blocks were aligned by ClustalW and refined manually.

In addition, these two variable regions contain full of indels and substitutions that can be explained by illegitimate recombination. For discussing this possibility, the boundaries of indels or substitutions were explored. In the following three cases, short repeats ranging from 3 to 9 bp were found, which indicates illegitimate recombination between these short repeats, as detailed below. First, in apparent VS2 deletion, between *rrnA* and *rrnF*, for example, 9-bp flanking direct repeats were found. Second, the apparent VS5-VS6 deletion, between A48073 and *rrnL*, for example, had 4-bp flanking direct repeats. Third, in substitution between VS9 and VS10, VS10 is apparently inserted with 3-bp flanking direct repeats instead of VS9. In addition, illegitimate recombination between 10-bp repeats can be inferred exclusively in MRSA252 for its *rrnH* and *rrnZ* due to the base substitutions and characteristic insertion of VS13. Their left and right variable regions thus generated by illegitimate recombination may recombine through homologous recombination at the conserved regions (CS1 and CS2).

Further Discussion

Most of the smaller polymorphisms were indels. They were deduced to be caused locally by illegitimate

Table 3
Types of 16S-23S rDNA Intergenic Spacer Region of the Seven Sequenced *Staphyloccous aureus* Strains

		Locus 2				
Strain	Locus 1	Locus 2-1	Locus 2-2	Locus 3	Locus 4	Locus 5
N315	*rrnC*	*rrnJ*		*rrnC*	*rrnE*	*rrnE*
Mu50	*rrnH*	*rrnJ*		*rrnC*	*rrnE*	*rrnE*
NCTC8325	*rrnH*	*rrnH*		*rrnC*	(A48073)	*rrnF*
COL	*rrnH*	*rrnJ*	*rrnL*	*rrnC*	(A48073)	(A48073)
MW2	*rrnL*	*rrnK*	*rrnJ*	*rrnA*	(A48073)	***rrnY***
MSSA476	*rrnL*	*rrnJ*	*rrnK*	*rrnA*	(A48073)	***rrnY***
MRSA252	***rrnY***	*rrnF*		*rrnC*	*rrnH*	***rrnZ***

NOTE.—The types (alleles) of 16S-23S rDNA intergenic spacer region are represented by names as described (Gurtler and Barrie 1995) except for A48073, named after the GenBank accession number, and for the two novel alleles, *rrnY* and *rrn Z* (shown in bold), identified in this study. See also figure 6.

recombination between flanking direct repeats or between sequences with no or very little similarity, though, in two cases, involvement of homologous recombination cannot be ruled out. Deletion appeared more likely than insertion between a pair of dispersed repeats (fig. 1*A* and *B* and table 2*A* and *B*), but once sequences were repeated in multiple times, they might become easy to expand and contract (fig. 1*C* and table 2*C*). Some of these sites may be used for markers for strain typing and may provide a tool to trace the route of infection (Read et al. 2002).

In all the three tandem paralogue clusters on νSaα and νSaβ, highly similar repeats, sufficiently long for homologous recombination, were identified at the boundaries of genome rearrangements. The relative positions of these repeats to the ORFs are apparently the same, so these repeats were conserved parts among these divergent tandem gene clusters. Homologous recombination between these repeats may have caused the interstrain copy number variation of the tandem paralogous genes. We cannot exclude the possibility of site-specific recombination for these and the other indels (blue indel and green indel in fig. 2*A* and *B*).

Extent of rearrangements appears to vary from one cluster to another as seen in the variation of gene order; the gene order appears to be conserved in the *ssl* cluster but far from conserved in the *lpl* cluster (figs. 2 and 3). This difference may be explained by the extent of divergence of tandem genes in each cluster and the strong dependency of the frequency of homologous recombination on the homology length and the sequence identity. In the *lpl* cluster, the conserved repetitive regions with sufficient length and sequence identity for homologous recombination exist in many pairs of overall repetitive structure corresponding to the tandem genes, which makes characteristic lattice pattern of dots in the dot plot (fig. 3*A*). This abundance of repeats that potentially cause homologous recombination is well correlated with the extensive rearrangements of the *lpl* cluster, which possibly include insertions, deletions, duplications, conversions, and translocations. A similar situation is seen in the left-hand side of the *spl* cluster as shown in figure 4. On the other hand, there seems no such abundance of significant repeats in the *ssl* cluster, which is coherent with the presence of only few indels in this cluster. If the sequence conserved within the *lpl* genes served as the recombination site, one might expect that phylogeny of the gene could be different to its left and to its right. This is indeed the case (T. Tsuru and I. Kobayashi, unpublished data).

Can the maintenance of these conserved repeats be interpreted in terms of their function? One pair in the *ssl* cluster (red in fig. 2) and those in the *spl* cluster (fig. 4) encode the initiating ATG, the putative ribosome-binding site, and a part of the putative signal peptides (Williams et al. 2000) and could be important in gene expression and secretion. The *ssl* transcripts were detected in an *S. aureus* strain (Williams et al. 2000), and the *spl* transcripts were detected in a derivative of NCTC8325 (Reed et al. 2001). The repeats in *spl* genes include the 108-bp sequence for the signal peptides, which is cleaved during secretion (fig. 4*C*) (Reed et al. 2001). Meanwhile, those repeats in the *lpl* cluster (fig. 3) correspond to conserved amino acid sequences, yet, these conserved residues were outside the recognized sequence motif for lipoproteins (Prosite accession number, PS00013) (http://www.expasy.org/prosite/). Lastly, another sequence repeated in the *ssl* cluster (orange in fig. 2) is located in the center of an intergenic region. We found no evidence for spread of these repetitive regions outside these genomic islands.

Is there any meaning in this variability among strains? These three tandem paralogues (*ssl*, *lpl*, and *spl*) have been considered as virulence genes as they encode exotoxins, lipoproteins, and secreted proteases, respectively (Williams et al. 2000; Kuroda et al. 2001; Reed et al. 2001). That all the three kinds of proteins are likely involved in interaction with environments suggests that the intraspecies variability of these clusters may confer ability to adapt to diverse environmental conditions. Many bacteria adapt to variable environments by altering a phenotype through changes in expression of multiple genes based on DNA rearrangements and other means—a process called "phase variation" (van der Woude and Baumler 2004). Variability of these clusters may be explained in the same context. We do not know whether these repeats are maintained because of their function as recombination sites that control the variability of a paralogue cluster.

Genetic exchanges via horizontal gene transfer in this region have been proposed because this region is on a putative mobile genetic element (Fitzgerald et al. 2003). The apparently distorted gene order in the *lpl* cluster and the *spl* cluster supports this view. The polymorphic pattern of *hsdS* genes that shows significant sequence similarity between different strains is far more indicative; an interstrain reassortment of target recognition domains might have occurred.

There is evidence that RM genes have undergone extensive horizontal gene transfer (Kobayashi 2004). Many RM genes of type I and type II are present on various mobile genetic elements (Kobayashi 2004). In the case of type I RM genes, aberrant GC contents and presence of different alleles of the same subfamily support this view (Murray 2000). The significant sequence similarity between conserved domains of *hsdS* in *S. aureus* and those in *L. lactis* also indicates that the exchange of the *hsdS* genes was possible even across their interspecies barriers. Similarity was also present between their M subunits. The presence of type I RM genes on the genomic islands has lead us to hypothesize that these type I RM genes could play a role in stabilizing the maintenance of these islands (Kuroda et al. 2001) because of postsegregational host killing as reported for type II RM systems (Naito, Kusano, and Kobayashi 1995).

The type I RM genes on the two genomic islands, νSaα and νSaβ, encode only HsdS and HsdM subunits sufficient for the methylation activity but not HsdR subunit necessary for the endonuclease activity (Kuroda et al. 2001). This led us to hypothesize that the HsdR subunit encoded in a locus outside the two genomic islands can interact with two sets of HsdM and HsdS to form two different RM systems (Kuroda et al. 2001). Combination of this HsdR and HsdM/HsdS of νSaα may form a restriction enzyme of one specificity, while combination of this HsdR and HsdM/HsdS of νSaβ may form another restriction enzyme of another specificity. This hypothesis is worth testing for the following reasons. First, the sequences of conserved domains of HsdS on the two different islands are very similar to each other. These domains are responsible for interacting with HsdR (Abadjieva et al. 1993; MacWilliams and Bickle 1996; Kim et al. 2005). Second, their homologous systems in *L. lactis*, Lla7I, Lla103I, and Lla1403I, comprise chromosomally encoded HsdM, HsdR, and HsdS genes and plasmid-encoded HsdS genes to confer combinatorial variation of restriction specificity by switching the HsdS subunit (Schouler et al. 1998).

With respect to the *rrn* operons, occurrence of homologous recombination involving long homology of rDNAs was inferred between two tandem loci, locus 2-1 and locus 2-2. This process was inferred to be deletion rather than duplication from detailed comparison of intergenic sequences. We do not know why the duplicated version (locus 2-1 and locus 2-2) does not segregate into a monomer version (locus 2) by homologous recombination. The selective advantage of the former in ribosomal function could provide an explanation. Manifestation of the homologous recombination between distant loci is also gained by the typing of 16S-23S rDNA spacer.

The presence of short repeats in the 16S-23S rDNA spacer rearrangements indicated involvement of illegitimate recombination between short repeats in the formation of its novel allele. Among their examples, the 9-bp repeats, which involved in VS2 deletion and consist of the 3′-end sequence of tRNA (Ala), CCACCA and following TTA, are noteworthy because the DNA secondary structure of the VS2 region which encodes tRNA (Ala) may have induced the illegitimate recombination. Indeed, computer analysis suggested that a stable secondary structure could be formed at tRNA (Ala) in *Hemophilus parainfluzae* (Giannino et al. 2001), and presence of such secondary structure stimulates illegitimate recombination (Michel 1999).

We have focused on polymorphisms that exist between N315 and Mu50 because we were involved in their genome analysis. Their genome sequences were close enough to allow reconstruction of their formation. A comparison of all the seven genomes would very likely identify additional and significant polymorphisms. However, how these polymorphisms were formed would be quite difficult to analyze with diverged genomes. There are more than six additional pathogenicity islands (excluding bacteriophages) in the N315/Mu50 genome and the remaining five genomes that also likely contribute to virulence. A truly global comparison of *S. aureus* genomes would include an analysis of the entire complement of pathogenicity islands.

After the manuscript was prepared, two more genome sequences of *S. aureus* strains, RF122 (RefSeq: NC_007622) and USA300 (RefSeq: NC_007793), were released.

Conclusion

In the present work, we compared multiple genome sequences of *S. aureus* to deduce mechanisms of genome rearrangements, in paralogous genes and others, that resulted in large genome polymorphisms. Most of them were inferred to have resulted from deletion through illegitimate recombination. For the tandem paralogue gene clusters on genomic islands, νSaα and νSaβ, we were able to identify sequences for homologous recombination. The evolution through homologous recombination was also found for the type I RM *hsdS* genes on these islands. Homologous recombination likely has caused the rearrangements in the rRNA operons. We also found novel alleles in rRNA operons and suggested involvement of illegitimate recombination in their formation. Taken together, these results demonstrate the importance of homologous recombination in the evolution of paralogous genes and illustrate power of comparative genomics in the analysis of genome evolution through genome rearrangements.

Acknowledgments

We are grateful to Toshiko Ohta and Hideki Hirakawa for communication of unpublished results and to Makoto Kuroda and Harumi Yuzawa for helpful comments on the manuscript. This work was supported by grants from Ministry of Education, Culture, Sports, Science, and Technology of the Japanese government to I.K. (Genome Biology, Genome Homeostasis, DNA Repair, Kiban-evolution, Kiban-genome, 21COE: genome language).

Literature Cited

Abadjieva, A., J. Patel, M. Webb, V. Zinkevich, and K. Firman. 1993. A deletion mutant of the type IC restriction endonuclease EcoR1241 expressing a novel DNA specificity. Nucleic Acids Res. **21**:4435–4443.

Adamczyk-Poplawska, M., A. Kondrzycka, K. Urbanek, and A. Piekarowicz. 2003. Tetra-amino-acid tandem repeats are involved in HsdS complementation in type IC restriction-modification systems. Microbiology **149**:3311–3319.

Alm, R. A., L. S. Ling, D. T. Moir et al. (26 co-authors). 1999. Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. Nature **397**: 176–180.

Baba, T., F. Takeuchi, M. Kuroda et al. (17 co-authors). 2002. Genome and virulence determinants of high virulence community-acquired MRSA. Lancet **359**:1819–1827.

Baba, T., F. Takeuchi, M. Kuroda, H. Yuzawa, T. Ito, and K. Hiramatsu. 2004. The genome of *Staphylococcus aureus*. Pp. 66–153 *in* D. A. A. Ala'Aladeen and K. Hiramatsu, eds. The *Staphylococcus aureus*: molecular and clinical aspects. Horwood, Chichester, United Kingdom.

Brussow, H., C. Canchaya, and W. D. Hardt. 2004. Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. Microbiol. Mol. Biol. Rev. **68**:560–602.

Chedin, F., E. Dervyn, R. Dervyn, S. D. Ehrlich, and P. Noirot. 1994. Frequency of deletion formation decreases exponentially with distance between short direct repeats. Mol. Microbiol. **12**:561–569.

Chin, V., V. Valinluck, S. Magaki, and J. Ryu. 2004. KpnBI is the prototype of a new family (IE) of bacterial type I restriction-modification system. Nucleic Acids Res. **32**:e138.

Datta, A., M. Hendrix, M. Lipsitch, and S. Jinks-Robertson. 1997. Dual roles for DNA sequence identity and the mismatch repair system in the regulation of mitotic crossing-over in yeast. Proc. Natl. Acad. Sci. USA **94**:9757–9762.

Fitzgerald, J. R., S. D. Reid, E. Ruotsalainen, T. J. Tripp, M. Liu, R. Cole, P. Kuusela, P. M. Schlievert, A. Jarvinen, and J. M. Musser. 2003. Genome diversification in *Staphylococcus aureus*: molecular evolution of a highly variable chromosomal region encoding the staphylococcal exotoxin-like family of proteins. Infect. Immun. **71**:2827–2838.

Forsman, P., A. Tilsala-Timisjarvi, and T. Alatossava. 1997. Identification of staphylococcal and streptococcal causes of bovine mastitis using 16S-23S rRNA spacer regions. Microbiology **143**:3491–3500.

Foster, T. J., and M. Hook. 1998. Surface protein adhesins of *Staphylococcus aureus*. Trends Microbiol. **6**:484–488.

Fujitani, Y., and I. Kobayashi. 1999. Effect of DNA sequence divergence on homologous recombination as analyzed by a random-walk model. Genetics **153**:1973–1988.

Fujitani, Y., K. Yamamoto, and I. Kobayashi. 1995. Dependence of frequency of homologous recombination on the homology length. Genetics **140**:797–809.

Fuller-Pace, F. V., L. R. Bullas, H. Delius, and N. E. Murray. 1984. Genetic recombination can generate altered restriction specificity. Proc. Natl. Acad. Sci. USA **81**:6095–6099.

Gann, A. A., A. J. Campbell, J. F. Collins, A. F. Coulson, and N. E. Murray. 1987. Reassortment of DNA recognition domains and the evolution of new specificities. Mol. Microbiol. **1**:13–22.

Giannino, V., G. Rappazzo, A. Scuto, O. Di Marco, A. Privitera, M. Santagati, and S. Stefani. 2001. rrn operons in *Haemophilus parainfluenzae* and mosaicism of conserved and species-specific sequences in the 16S-23S rDNA long spacer. Res. Microbiol. **152**:461–468.

Gill, S. R., D. E. Fouts, G. L. Archer et al. (32 co-authors). 2005. Insights on evolution of virulence and resistance from the complete genome analysis of an early methicillin-resistant *Staphylococcus aureus* strain and a biofilm-producing methicillin-resistant *Staphylococcus epidermidis* strain. J. Bacteriol. **187**:2426–2438.

Gubler, M., D. Braguglia, J. Meyer, A. Piekarowicz, and T. A. Bickle. 1992. Recombination of constant and variable modules alters DNA sequence recognition by type IC restriction-modification enzymes. Embo J. **11**:233–240.

Gurtler, V. 1999. The role of recombination and mutation in 16S-23S rDNA spacer rearrangements. Gene **238**:241–252.

Gurtler, V., and H. D. Barrie. 1995. Typing of *Staphylococcus aureus* strains by PCR-amplification of variable-length 16S-23S rDNA spacer regions: characterization of spacer sequences. Microbiology **141**:1255–1265.

Gurtler, V., and B. C. Mayall. 2001. Genetic transfer and genome evolution in MRSA. Microbiology **147**:3195–3197.

Gurtler, V., and V. A. Stanisich. 1996. New approaches to typing and identification of bacteria using the 16S-23S rDNA spacer region. Microbiology **142**:3–16.

Harvey, S., C. W. Hill, C. Squires, and C. L. Squires. 1988. Loss of the spacer loop sequence from the *rrnB* operon in the *Escherichia coli* K-12 subline that bears the *relA1* mutation. J. Bacteriol. **170**:1235–1238.

Hill, C. W. 1999. Large genomic sequence repetitions in bacteria: lessons from rRNA operons and Rhs elements. Res. Microbiol. **150**:665–674.

Hiramatsu, K., L. Cui, M. Kuroda, and T. Ito. 2001. The emergence and evolution of methicillin-resistant *Staphylococcus aureus*. Trends Microbiol. **9**:486–493.

Holden, M. T., E. J. Feil, J. A. Lindsay et al. (48 co-authors). 2004. Complete genomes of two clinical *Staphylococcus aureus* strains: evidence for the rapid evolution of virulence and drug resistance. Proc. Natl. Acad. Sci. USA **101**: 9786–9791.

Katayama, Y., T. Ito, and K. Hiramatsu. 2000. A new class of genetic element, staphylococcus cassette chromosome mec, encodes methicillin resistance in *Staphylococcus aureus*. Antimicrob. Agents Chemother. **44**:1549–1555.

Khasanov, F. K., D. J. Zvingila, A. A. Zainullin, A. A. Prozorov, and V. I. Bashkirov. 1992. Homologous recombination between plasmid and chromosomal DNA in *Bacillus subtilis* requires approximately 70 bp of homology. Mol. Gen. Genet. **234**:494–497.

Kim, J. S., A. DeGiovanni, J. Jancarik, P. D. Adams, H. Yokota, R. Kim, and S. H. Kim. 2005. Crystal structure of DNA sequence specificity subunit of a type I restriction-modification enzyme and its functional implications. Proc. Natl. Acad. Sci. USA **102**:3248–3253.

Kobayashi, I. 2004. Restriction-modification systems as minimal forms of life. Pp. 19–62 *in* A. Pingoud, ed. Restriction endonucleases. Springer-Verlag, Berlin, Germany.

Kuroda, M., T. Ohta, I. Uchiyama et al. (37 co-authors). 2001. Whole genome sequencing of meticillin-resistant *Staphylococcus aureus*. Lancet **357**:1225–1240.

Lan, R., and P. R. Reeves. 1998. Recombination between rRNA operons created most of the ribotype variation observed in the seventh pandemic clone of *Vibrio cholerae*. Microbiology **144**:1213–1221.

Liao, D. 2000. Gene conversion drives within genic sequences: concerted evolution of ribosomal RNA genes in bacteria and archaea. J. Mol. Evol. **51**:305–317.

Lina, G., G. A. Bohach, S. P. Nair, K. Hiramatsu, E. Jouvin-Marche, and R. Mariuzza. 2004. Standard nomenclature for the superantigens expressed by *Staphylococcus*. J. Infect. Dis. **189**:2334–2336.

Lindsay, J. A., and M. T. Holden. 2004. *Staphylococcus aureus*: superbug, super genome? Trends Microbiol. **12**: 378–385.

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

Lovett, S. T., T. J. Gluckman, P. J. Simon, V. A. Sutera Jr, and P. T. Drapkin. 1994. Recombination between repeats in *Escherichia coli* by a *recA*-independent, proximity-sensitive mechanism. Mol. Gen. Genet. **245**:294–300.

MacWilliams, M. P., and T. A. Bickle. 1996. Generation of new DNA binding specificity by truncation of the type IC EcoDXXI *hsdS* gene. Embo J. **15**:4775–4783.

McDevitt, D., and T. J. Foster. 1995. Variation in the size of the repeat region of the fibrinogen receptor (clumping factor) of *Staphylococcus aureus* strains. Microbiology **141**: 937–943.

Michel, B. 1999. Illegitimate recombination in bacteria. Pp. 129–150 *in* R. L. Charlebois, ed. Organization of the prokaryotic genome. ASM Press, Washington, D. C.

Murray, N. E. 2000. Type I restriction systems: sophisticated molecular machines (a legacy of Bertani and Weigle). Microbiol. Mol. Biol. Rev. **64**:412–434.

Naito, T., K. Kusano, and I. Kobayashi. 1995. Selfish behavior of restriction-modification systems. Science **267**:897–899.

Nei, M., and S. Kumar. 2001. Molecular evolution and phylogenetics. Oxford University Press, New York.

Nobusato, A., I. Uchiyama, S. Ohashi, and I. Kobayashi. 2000. Insertion with long target duplication: a mechanism for gene mobility suggested from comparison of two related bacterial genomes. Gene **259**:99–108.

Ohta, T., H. Hirakawa, K. Morikawa et al. (12 co-authors). 2004. Nucleotide substitutions in *Staphylococcus aureus* strains, Mu50, Mu3, and N315. DNA Res. **11**:51–56.

Privitera, A., G. Rappazzo, P. Sangari, V. Giannino, L. Licciardello, and S. Stefani. 1998. Cloning and sequencing of a 16S/23S ribosomal spacer from *Haemophilus parainfluenzae* reveals an invariant, mosaic-like organisation of sequence blocks. FEMS Microbiol. Lett. **164**:289–294.

Read, T. D., S. L. Salzberg, M. Pop et al. (13 co-authors). 2002. Comparative genome sequencing for discovery of novel polymorphisms in *Bacillus anthracis*. Science **296**:2028–2033.

Reed, S. B., C. A. Wesson, L. E. Liou, W. R. Trumble, P. M. Schlievert, G. A. Bohach, and K. W. Bayles. 2001. Molecular characterization of a novel *Staphylococcus aureus* serine protease operon. Infect. Immun. **69**:1521–1527.

Robinson, D. A., and M. C. Enright. 2004. Evolution of *Staphylococcus aureus* by large chromosomal replacements. J. Bacteriol. **186**:1060–1064.

Rocha, E. P. 2004. Order and disorder in bacterial genomes. Curr. Opin. Microbiol. **7**:519–527.

Rocha, E. P., and A. Blanchard. 2002. Genomic repeats, genome plasticity and the dynamics of *Mycoplasma* evolution. Nucleic Acids Res. **30**:2031–2042.

Ruzin, A., J. Lindsay, and R. P. Novick. 2001. Molecular genetics of SaPI1—a mobile pathogenicity island in *Staphylococcus aureus*. Mol. Microbiol. **41**:365–377.

Schouler, C., M. Gautier, S. D. Ehrlich, and M. C. Chopin. 1998. Combinational variation of restriction modification specificities in *Lactococcus lactis*. Mol. Microbiol. **28**:169–178.

Titheradge, A. J., J. King, J. Ryu, and N. E. Murray. 2001. Families of restriction enzymes: an analysis prompted by molecular and genetic data for type ID restriction and modification systems. Nucleic Acids Res. **29**:4195–4205.

Uchiyama, I., T. Higuchi, and I. Kobayashi. 2000. CGAT: comparative genome analysis tool for closely related microbial genomes. Genome Informatics **11**:341–342.

van der Woude, M. W., and A. J. Baumler. 2004. Phase and antigenic variation in bacteria. Clin. Microbiol. Rev. **17**:581–611.

Vulic, M., F. Dionisio, F. Taddei, and M. Radman. 1997. Molecular keys to speciation: DNA polymorphism and the control of genetic exchange in enterobacteria. Proc. Natl. Acad. Sci. USA **94**:9763–9767.

Williams, R. J., J. M. Ward, B. Henderson, S. Poole, B. P. O'Hara, M. Wilson, and S. P. Nair. 2000. Identification of a novel gene cluster encoding staphylococcal exotoxin-like proteins: characterization of the prototypic gene and its protein product, SET1. Infect. Immun. **68**:4407–4415.

Yamaguchi, T., K. Nishifuji, M. Sasaki, Y. Fudaba, M. Aepfelbacher, T. Takata, M. Ohara, H. Komatsuzawa, M. Amagai, and M. Sugai. 2002. Identification of the *Staphylococcus aureus etd* pathogenicity island which encodes a novel exfoliative toxin, ETD, and EDIN-B. Infect. Immun. **70**:5835–5845.

Takashi Gojoboroi, Associate Editor

Accepted March 29, 2006