Multiple Genome Comparison within a Bacterial Species Reveals a Unit of Evolution Spanning Two Adjacent Genes in a Tandem Paralog Cluster

Takeshi Tsuru†‡ and Ichizo Kobayashi*†‡*

*Laboratory of Social Genome Sciences, Department of Medical Genome Sciences, Graduate School of Frontier Sciences, University of Tokyo, Tokyo, Japan; †Graduate Program in Biophysics and Biochemistry, Graduate School of Science, University of Tokyo, Tokyo, Japan; and ‡Institute of Medical Science, University of Tokyo, Tokyo, Japan

It has been assumed that an open reading frame (ORF) represents a unit of gene evolution as well as a unit of gene expression and function. In the present work, we report a case in which a unit comprising the 3′ region of an ORF linked to a downstream intergenic region that is in turn linked to the 5′ region of a downstream ORF has been conserved, and has served as the unit of gene evolution. The genes are tandem paralogous genes from the bacterium *Staphylococcus aureus*, for which more than ten entire genomes have been sequenced. We compared these multiple genome sequences at a locus for the *lpl* (*l*ipo*p*rotein-*l*ike) cluster (encoding lipoprotein homologs presumably related to their host interaction) in the genomic island termed νSaα. A highly conserved nucleotide sequence found within every *lpl* ORF is likely to provide a site for homologous recombination. Comparison of phylogenies of the 5′-variable region and the 3′-variable region within the same ORF revealed significant incongruence. In contrast, pairs of the 3′-variable region of an ORF and the 5′-variable region of the next downstream ORF gave more congruent phylogenies, with distinct groups of conserved pairs. The intergenic region seemed to have coevolved with the flanking variable regions. Multiple recombination events at the central conserved region appear to have caused various types of rearrangements among strains, shuffling the two variable regions in one ORF, but maintaining a conserved unit comprising the 3′-variable region, the intergenic region, and the 5′-variable region spanning adjacent ORFs. This result has strong impact on our understanding of gene evolution because most gene lineages underwent tandem duplication and then diversified. This work also illustrates the use of multiple genome sequences for high-resolution evolutionary analysis within the same species.

Introduction

Gene duplication has long been recognized as an important mechanism in the evolution of genes. Since the essential role of gene duplication in the emergence of novel genes was proposed (Ohno 1970), numerous works have studied the mechanisms by which the duplicated genes diversified (Li 1997). Recent progress in genome sequencing has confirmed the significance of gene duplication by revealing a wide prevalence of multiple homologous genes within a genome, often referred to simply as paralogs (Lynch and Conery 2000; Friedman and Hughes 2001; Gevers et al. 2004). Comparison of multiple genome sequences has provided ample evidence that the present paralogs have undergone various patterns of diversification (Prince and Pickett 2002; Taylor and Raes 2004).

In bacteria, many paralogous gene groups have been shown to be involved in genome rearrangements that help the bacteria adapt to ever-changing environments. These processes are referred to as phase variation or antigenic variation, and the dynamics of paralog evolution in bacteria have been mostly examined with reference to these mechanisms (van der Woude and Baumler 2004; Villemur and Deziel 2005). Studies have revealed that these genes have become rearranged through simple processes such as inversion, deletion, or gene conversion, via various molecular mechanisms such as site-specific recombination, homologous recombination, or slipped-strand mispairing during DNA replication (Hughes and Norstrom 2005; Villemur and Deziel 2005). Many paralogous genes that have undergone these rearrangements have been discovered through whole-genome sequencing. Comparison of closely related prokaryotic genomes can help in the elucidation of the molecular mechanisms underlying the rearrangements of paralogous genes.

Paralogous genes are often present as a tandem cluster in prokaryotic genomes (Gevers et al. 2004; Reams and Neidle 2004). They often encode surface proteins (Kihara and Kanehisa 2000; Gevers et al. 2004), which may undergo phase variation or antigenic variation (Hughes and Norstrom 2005; Villemur and Deziel 2005). Because their genes were likely to have originated from duplication, tandem paralogs can be suitable targets for the study of paralog diversification. Some of these paralogous gene clusters are on genomic islands, which are likely to have been horizontally acquired, to be highly polymorphic among strains, and to confer strain-specific adaptive properties such as drug resistance or pathogenicity (Dobrindt et al. 2004).

Considering the above point, tandem paralogs found in the genomic islands of *Staphylococcus aureus* are suitable targets for the study of paralog evolution. *Staphylococcus aureus* is a Gram-positive bacterium with low GC content. It is a major human pathogen and is notable for its expression of a vast variety of toxins (Kuroda et al. 2001). Whole-genome sequences have been determined for more than 10 *S. aureus* strains (Kuroda et al. 2001; Baba et al. 2002, 2008; Holden et al. 2004; Ohta et al. 2004; Gill et al. 2005; Diep et al. 2006; Highlander et al. 2007; http://genome.jgi-psf.org/finished_microbes/). Two genomic islands, νSaα and νSaβ, common and unique to this bacterial species, encode three tandem paralog clusters: exotoxins (*ssl*) and lipoproteins (*l*ipo*p*rotein-*l*ike [*lpl*]) encoded by νSaα and proteases (*spl*) encoded by νSaβ. All these paralogs encode secreted proteins that are inferred to be pathogenicity related (Williams et al. 2000; Kuroda et al. 2001; Reed et al. 2001; Chavakis et al. 2007; Kulig et al. 2007). The copy number and sequence composition varies among strains. In our previous study (Tsuru et al. 2006), we compared these three clusters from seven

Key words: *Staphylococcus aureus*, genome evolution, gene duplication, genome rearrangement, genome comparison, MRSA.

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

Mol. Biol. Evol. 25(11):2457–2473. 2008
doi:10.1093/molbev/msn192
Advance Access publication September 2, 2008

strains available at the time and inferred an involvement of homologous recombination in their evolution. Among the three, the *lpl* cluster was unique in that highly conserved nucleotide sequences, representing a possible recombination site, were discovered at the middle of the protein-coding region of every paralogous open reading frame (ORF) and in that rearrangements there seemed to be far more extensive than in the other two clusters.

In the present study, we made a detailed comparison of the *lpl* tandem paralog clusters of nine sequenced *S. aureus* genomes to examine their evolutionary processes. Contrary to the general belief that a unit of evolution of a gene is an ORF itself, we came to the conclusion that a unit composed of the "3′ half of an ORF and the 5′ half of a downstream ORF" serves as a unit of evolution in this cluster.

Materials and Methods

Homology Search for *lpl* Genes

The annotated *lpl* amino acid sequences of *S. aureus* N315 were used as the query sequences to search for homologs of *lpl* genes using BlastP and TBlastN. The search was carried out against the complete genome sequences of 520 bacterial and 46 archaeal species (31 July 2007 data) from the *National Center for Biotechnology Information* (NCBI) Genome database (http://www.ncbi.nlm.nih.gov/Genomes/). All hits with an *e* value $< 10^{-5}$ were collected. Pseudogenes of *lpl* were identified through these analyses, some of which had been referred as such in the original annotations. Against the *lpl* homologs, manual refinement at the nucleotide sequence level was carried out by macroscopic pairwise genome comparison using CGAT software (Uchiyama et al. 2006) and by multiple sequence alignment using ClustalW version 1.83 (http://www.ebi.ac.uk/Tools/clustalw/). The initiation and termination positions for ORFs (including pseudogenes) were reassigned consistently to resolve discrepancy between the original annotators. The resulting coordinates for all the *lpl* ORFs are listed in supplementary table S1 (Supplementary Material online).

Nomenclature of Genes

Names for ORFs and pseudogenes used throughout this study were set by modifying the corresponding locus_tag in RefSeq database (http://www.ncbi.nlm.nih.gov/RefSeq/): a three-letter genome name used in the Kyoto Encyclopedia of Genes and Genomes database (http://www.genome.jp/kegg/) was followed by the locus_tag number, for example, SAA0410 for SAUSA300_0410 in RefSeq. The strains in which the relevant homologs were identified and their three-letter genome names are as follows. In *S. aureus*, SAU is for strain N315 (Kuroda et al. 2001), SAV for Mu50 (Kuroda et al. 2001; Ohta et al. 2004), SAM for MW2 (Baba et al. 2002), SAR for MRSA252 (Holden et al. 2004), SAS for MSSA476 (Holden et al. 2004), SAC for COL (Gill et al. 2005), SAO for NCTC8325 (http://microgen.ouhsc.edu/s_aureus/s_aureus_home.htm), SAA for USA300 (Diep et al. 2006), and SAB for RF122 (Herron-Olson et al. 2007). In *Staphylococcus epidermidis*, SEP is for strain ATCC12228 (Zhang et al. 2003) and SER for RP62A (Gill et al. 2005). In *Staphylococcus haemolyticus*, SHA is for strain JCSC1435 (Takeuchi et al. 2005). A pseudogene was indicated by adding "p" at the end of the name, for example, SAA0412p.

Sequence Comparison

An initial multiple sequence alignment of the nucleotide sequences for all the *lpl* ORFs was constructed using ClustalW with the default parameters. Then a Neighbor-Joining (NJ) phylogeny for the alignment was constructed using MEGA 4.0 (Tamura et al. 2007; http://www.megasoftware.net/) with pairwise deletion mode for gaps and with a maximum composite likelihood model for substitutions (fig. 1*B* and supplementary fig. S1*A* [Supplementary Material online]).

More detailed analyses were carried out with *lpl* ORFs on genomic island νSaα of *S. aureus* as detailed below. For the nucleotide sequences of the ORFs, multiple sequence alignment was once again constructed using ClustalW. A multiple sequence alignment for the predicted amino acid sequences was also constructed, omitting those for the pseudogenes. The nucleotide sequence alignment was then manually refined considering their encoding amino acid sequences, using a function implemented in MEGA 4.0. The resulting nucleotide and amino acid sequence alignments are shown in supplementary figure S3*A* and *B* (Supplementary Material online), respectively. The nucleotide sequences of the relevant regions including the ORFs and the intergenic regions are identical between N315 and Mu50 and between MW2 and MSSA476; therefore, the sequences of Mu50 and MSSA476 were omitted here and in the following analyses.

Definitions Related to the Structure of *lpl* ORFs on νSaα

The presence of the central conserved region was visualized by similarity plots of aligned nucleotide/amino acid sequences constructed using PLOTCON (http://emboss.sourceforge.net/), in which a similarity score, with window size of 5, was calculated with EDNAFULL score file for nucleotide sequences and with EBLOSUM62 for amino acid sequences (fig. 2). A region conserved both at the nucleotide sequence level and at the amino acid sequence level was determined in the alignments by visual inspection and defined as the central conserved region (supplementary fig. S3*A* and *B*, Supplementary Material online). Divergent regions to its 5′ side and to its 3′ side were defined as the 5′-variable region and the 3′-variable region, respectively.

Phylogenetic Comparison and Grouping of 5′-Variable Region and 3′-Variable Region

Nucleotide sequences of the 5′-variable region and the 3′-variable region were once again aligned using ClustalW, and NJ phylogenies for them were constructed using MEGA 4.0 with the complete deletion mode for gaps and with the maximum composite likelihood model for substitutions. The phylogenies were compared with each other by connecting operational taxonomic units (OTUs) in a pair

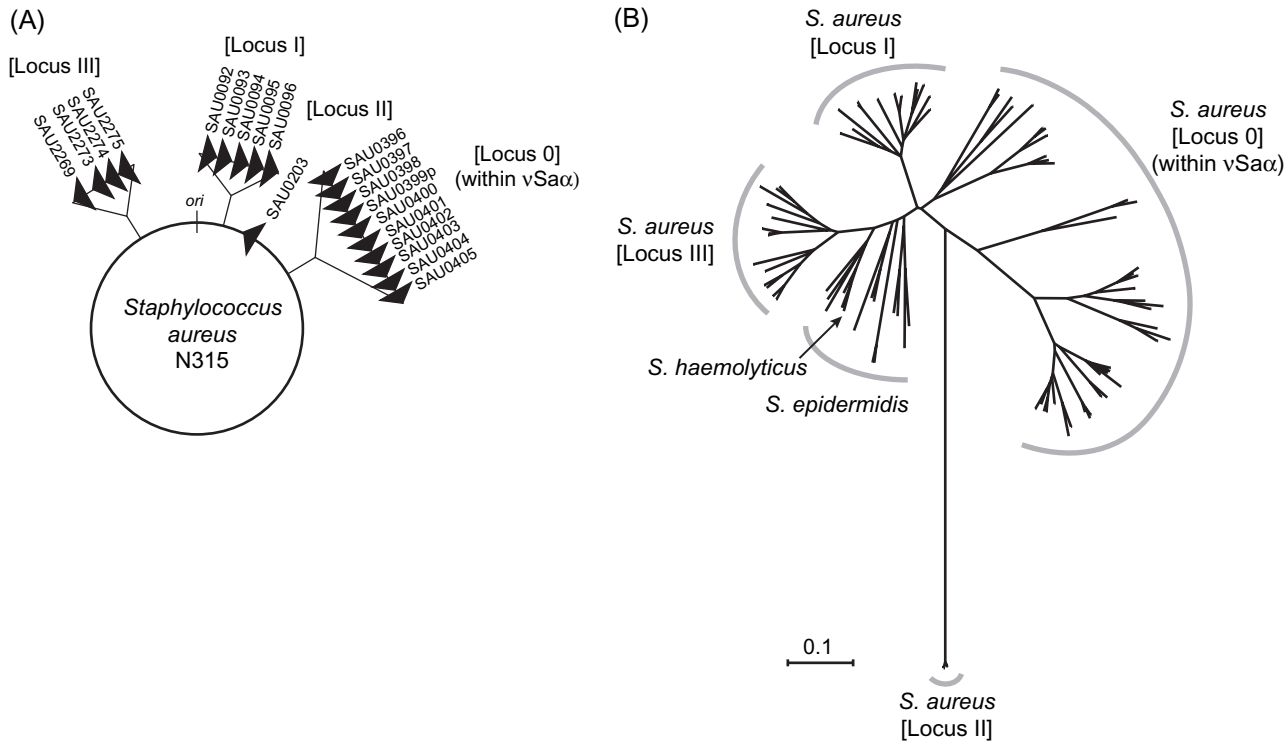

FIG. 1.—The *lpl* homologs in *Staphylococcus aureus* genome and their phylogenetic tree. (*A*) Location of four *lpl* loci, Locus 0 through Locus III, on the genome of strain N315. Note that the *lpl* homologs are found in the corresponding loci in all the sequenced *S. aureus* strains. (*B*) A nucleotide NJ phylogenetic tree for the *lpl* ORFs and their homologs in two other *Staphylococcus* species. The uncondensed version of this tree is presented in supplementary figure S1*A* (Supplementary Material online).

of the 3′-variable region of an ORF and the 5′-variable region of its downstream ORF (fig. 4*A*) and in a pair within an ORF (fig. 4*B*). Shapes of the trees were modified by flipping and rerooting to reduce the number of crosses of the connecting lines in each comparison with aid of TreeMap 2.0 (http://www.it.usyd.edu.au/~mcharles/software/treemap/treemap.html). Grouping was made for each variable region based on the comparison of phylogenetic trees in figure 4*A*. Pairwise identities with respect to the above grouping and with all the sequences were calculated using MEGA 4.0 (table 1).

For statistically testing whether our tree-based grouping reflects significant linkage between a 3′-variable region and its downstream 5′-variable region, we performed Fisher's exact test implemented in R packages version 2.7.0 (http://www.R-projects.org/), which is based on the FORTRAN program FEXACT (Mehta and Patel 1986; Clarkson et al. 1993).

Comparison of Intergenic Regions

Sequences of the intergenic regions were compared using ClustalW for each group. The presence of intergroup similarities was detected with aid of NJ trees using MEGA4.0 and multiple sequence alignments using ClustalW (fig. 5). Pairwise identities were calculated using MEGA4.0 (table 1). Ribosome-binding sites were predicted in the multiple alignments, referring to the characterized consensus sequences (Novick 1991).

Intergenomic Comparison

Intergenomic comparison was carried out using a map of ORFs shown in figure 6*A*. For indels, boundaries of homology and nonhomology were compared by ClustalW as described previously (Tsuru et al. 2006) to identify recombination sites (fig. 6*B* and *C*).

Analyses of Another Tandem Gene Cluster for Hypothetical Proteins in *S. aureus*

The above definitions related to gene structure and phylogenetic comparison were repeated for another tandem gene cluster for hypothetical proteins, homologs of SA1317 (figs. 8–10). Grouping was carried out for each of the two variable regions separately so as the mutual evolutionary distance remains equal to or shorter than 0.15 within a group (fig. 9).

Results

Distribution of *lpl* Homologues in *Staphylococcus* Genomes

In a previous study, we examined diversity of the *lpl* gene cluster on the genomic island νSaα of several *S. aureus* strains with a sequenced genome (Tsuru et al. 2006). It had been reported that the homologs of these *lpl* genes were found in other loci than the genomic island and that they compose the largest group of paralogous genes in *S. aureus* strain N315 (Kuroda

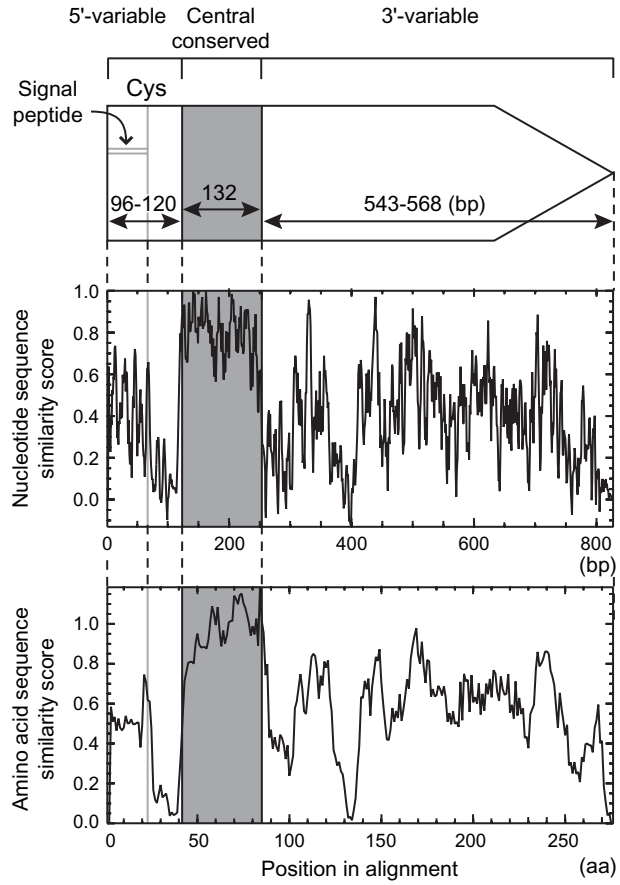

FIG. 2.—Structure of *lpl* genes (top) with similarity plots for nucleotide sequences (middle) and for amino acid sequences (bottom). The central conserved region is highlighted by gray shading. A predicted signal peptide region is indicated together with a conserved cysteine residue at the C-terminus.

et al. 2001; Baba et al. 2004). Therefore, we first carried out a search for their homologs against the microbial genome sequence database (NCBI Genome database; 31 July 2007 data, see Materials and Methods). This revealed that significantly related *lpl* homologs were found only in the genomes of the genus *Staphylococcus*. All the nine sequenced strains of *S. aureus* examined (N315, Mu50, MW2, MRSA252, MSSA476, COL, NCTC8325, USA300, and RF122) and two sequenced strains of *S. epidermidis* (ATCC12228 and RP62A) carry multiple *lpl* homologs within their genome, and the sequenced strain of *S. haemolyticus* (JCSC1435) carries one *lpl* homolog (supplementary table S1, Supplementary Material online). In the only remaining *Staphylococcus* genome that has been sequenced, *Staphylococcus saprophyticus* strain ATCC15305 (Kuroda et al. 2005), and in the other bacterial or archaeal genomes, no homolog could be found.

In all the *S. aureus* genomes, paralogous *lpl* genes were found in four loci (fig. 1*A*). First, in the genomic island νSaα, which we analyzed previously (Tsuru et al. 2006), the tandem cluster with three to ten *lpl* homologs is present in all nine strains (Locus 0). Second, one to five tandemly repeated *lpl* genes are found in a locus corresponding to SAU0092–SAU0096 of strain N315 in all the strains (Locus I). Third, only one *lpl* gene is present in a locus corresponding to SAU0203 of N315 in all the strains except for RF122 and MRSA252 (Locus II). Fourth, one to four tandemly repeated *lpl* genes, sometimes with intervening ORFs, can be found in a locus corresponding to SAU2269–SAU2275 of N315 in all the strains (Locus III). Note that all the nine *S. aureus* strains show synteny along the entire genome and carry the *lpl* genes at the same loci in the same orientation.

In order to examine orthologous/paralogous relationships among the *lpl* homologs, all the nucleotide sequences were compared in a multiple alignment to construct an NJ phylogeny. The condensed version of the resulting tree is displayed in figure 1*B* and the uncondensed version in supplementary figure S1*A* (Supplementary Material online). The tree revealed the presence of three monophyletic groups completely corresponding to Locus I, Locus II, and Locus III of *S. aureus*. Within each group, sequences are relatively similar to each other. Homologs from *S. epidermidis* and *S. haemolyticus* were rather divergent and were apparently related to the family at Locus III of *S. aureus*. The genes on νSaα of *S. aureus* (Locus 0) did not form a closely related monophyletic group, rather a mixture of several phylogenetic groups. However, these are clearly distinct from the groups at the other three loci of *S. aureus*. These observations suggested that the *lpl* genes of *S. aureus* have evolved, to the first approximation, separately at each locus (see also Discussion). Therefore, in the following sections, we will focus on sequence comparison among *lpl* homologs from the νSaα genomic island of *S. aureus*.

Central Conserved Region in *lpl* Genes on νSaα

In our previous study (Tsuru et al. 2006), we identified a conserved sequence within the *lpl* genes on νSaα through a multiple dot-plot analysis using the genome sequences of seven strains available at the time. We here confirmed the presence of this conserved sequence in two additional strains, USA300 and RF122, through the same analysis (supplementary fig. S2, Supplementary Material online). The presence of the conserved sequence is seen as dots at the crossing points of a horizontal red line and a vertical red line in comparison within the same genome, as noted in our previous work (Tsuru et al. 2006).

To verify this conserved sequence, multiple alignments for nucleotide sequences and for predicted amino acid sequences were constructed for all the relevant *lpl* ORFs (supplementary fig. S3*A* and *B*, Supplementary Material online). These alignments revealed the presence of a region conserved at both the nucleotide sequence level and the amino acid sequence level, which we named the central conserved region (fig. 2). This region is 132 nt long or 44 amino acids long without any gap in the alignments. The regions 5′ and 3′ of the central conserved region are less conserved and have variable lengths, and, accordingly, they were defined as the 5′-variable region and the 3′-variable region, respectively.

Calculation of pairwise nucleotide sequence identities using all the relevant sequences (table 1, All) revealed that those for the central conserved region are high (minimum of 79% and an average of 88%). Out of the 1,128 pairwise

Table 1
Statistics of Nucleotide Sequence Alignments for *lpl* ORFs and Their Intergenic Region

Name	Number of Sequences	Length in Alignment (bp)	Pairwise Identity	
			Minimum–Maximum (%)	Average (%)
5′-variable region				
All	48	120	41–100	65
a	6	96	89–100	94
b	3	96	94–98	96
c	3	87	100	100
d	2	96	NR	93
e	3	108	99–100	99
f	3	108	92–96	94
g	1	108	NR	NR
h	2	108	NR	98
i	2	108	NR	93
j	2	108	NR	95
k	11	96	80–100	88
l	1	96	NR	NR
m	9	120	90–100	96
Central conserved region				
All	48	132	79–100	88
3′-variable region				
All	48	574	52–100	66
A	8	561	87–100	94
B1	3	543	98–100	99
B2	2	543	NR	99
C	2	561	NR	98
D	3	546	98–100	99
E	3	561	91–92	92
F	3	565	88–93	91
G	6	561	91–100	94
H	2	557	NR	99
I	1	567	NR	NR
J	3	552	99–100	99
K	2	558	NR	89
L	1	552	NR	NR
M	9	567	82–100	90
Intergenic region				
A–a	6	18	100	100
B1–e	3	31	100	100
B2–h	2	31	NR	97
C–g	1	18	NR	NR
C–k	1	18	NR	NR
D–f	3	30	97–100	98
E–i	2	51	NR	98
F–b	3	47	89–98	93
G–k	3	47	92–96	94
H–j	2	59	NR	100
J–c	3	69	97–100	98
K–d	2	28	NR	93
L–l	1	60	NR	NR
M–m	9	223	81–100	92

NOTE.—NR, not relevant.

relationships, 427 showed identity as high as 90%. Meanwhile, those for the 5′-variable region and the 3′-variable region show minimal values of 41% and 52%, respectively, whereas the averages are 65% and 66%, respectively.

Involvement of this conserved sequence in the rearrangement of the *lpl* cluster is visible in an intergenomic dot-plot comparison (supplementary fig. S2, Supplementary Material online) as the termination of long black lines at a crossover point of a horizontal red line and a vertical red line as noted earlier (Tsuru et al. 2006). Generally, homologous recombination requires two homologous sequences long enough and similar enough to each other. In *Bacillus subtilis*, the closest bacterium to *S. aureus* in which this process has been studied in detail, the minimal length is 70 bp (Khasanov et al. 1992), a size comparable to those reported for other prokaryotic systems (Shen and Huang 1986; Fujitani et al. 1995). Frequency of homologous recombination is very sensitive to homology length around this length: it was found to be proportional to the third power of the homology length (Fujitani et al. 1995). The frequency of homologous recombination decreases very rapidly as the two sequences diverge (Vulic et al. 1997; Majewski and Cohan 1998; Fujitani and Kobayashi 1999). Thus, the homologous recombination between the central conserved regions is likely to occur.

Phylogeny Comparison and Grouping of 5′-Variable Region and 3′-Variable Region

If the central conserved region served as a recombination site during diversification of this region, the crossing-over events there should have changed the combination of the 5′-variable region and the 3′-variable region of an ORF (fig. 3). These events would result in incongruent phylogenies between them within an ORF. On the other hand, these crossing-over events will not disturb the linkage between 3′-variable region of an ORF and 5′-variable region of the next downstream ORF (fig. 3). They would result in a congruent phylogeny between these combinations. To test these predictions, a phylogeny for 5′-variable region and that for 3′-variable region were constructed and compared (fig. 4).

In figure 4*B* for a pair of variable regions within an ORF, the two phylogenies were found to be significantly incongruent. In figure 4*A* for a pair encompassing adjacent ORFs, the phylogenies appeared to be more congruent to each other. We observed that the connecting lines in figure 4*A* could be divided into groups of parallel lines with only a few exceptions. By contrast, in figure 4*B*, such clear groupings of the parallel connecting lines were difficult to identify (see Materials and Methods).

We also observed that the grouping of the parallel lines in figure 4*A* coincided with monophyletic or paraphyletic grouping of each phylogeny. Based upon this observation, we were able to assign groups for the 5′-variable region and the 3′-variable region, respectively. The 3′-variable regions were grouped into 13 distinct monophyletic groups named “A” through “M,” whereas the 5′-variable regions were also grouped into 13 distinct monophyletic groups named “a” through “m.” The group “B” was further divided into a paraphyletic group “B1” and a monophyletic group “B2” because these two are paired with distinct groups of the 5′-variable region in figure 4*A*: group B1 is paired with group “e” and group B2 is paired with “h,” respectively. The internal branch lengths for the resulting groups were relatively short, which indicated sequences are similar to each other within each group. Indeed, pairwise nucleotide identities calculated for each of the resulting groups were calculated to be $\geq$80%, which is in contrast to those calculated using all the sequences (table 1).

In all, 12 distinct groups of the pairs could be assigned in the combination of the 3′-variable region of an ORF and the 5′-variable region of its downstream ORF, which are

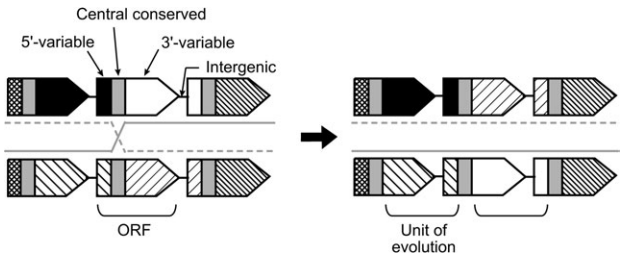

FIG. 3.—An elementary process of diversification through homologous recombination between the central conserved regions. A crossing-over will change combinations of the 5′-variable region and the 3′-variable region of an ORF, but linkage of 3′-variable region of an ORF, its downstream intergenic region, and 5′-variable region of its downstream ORF will be maintained.

displayed in different colors of connecting lines in figure 4*A*. The corresponding monophyletic or paraphyletic groups in each phylogeny are displayed in the same, but faint, colors as the connecting lines. One pair, "L" and "l," forms a group of only one member. The other groups have multiple members. The sequences of two variable regions are conserved within each group but are diverged between groups. There are only two exceptional pairs in this grouping, which are indicated by gray connecting lines in figure 4*A*. These are discussed in detail later (see Discussion).

Our grouping is based on comparison of the two phylogenetic trees in figure 4*A*. In order to examine whether it represents a statistically significant relationship, we performed Fisher's exact test (Agresti 1992; see Materials and Methods). The probability calculated for our data is 2.2×10^{-16}, which indicates that there is a highly significant linkage between a 3′-variable region and its downstream 5′-variable region.

This presence/absence of the conserved pairs in figure 4*A* and *B* supports the hypothesis that the two variable regions of one ORF have been shuffled during the diversification processes via crossing-over events at the central conserved region, whereas the two variable regions encompassing adjacent ORFs have been conserved during these processes.

Comparison of Intergenic Regions

If the linkage between the 3′-variable region of an ORF and the 5′-variable region of its downstream ORF

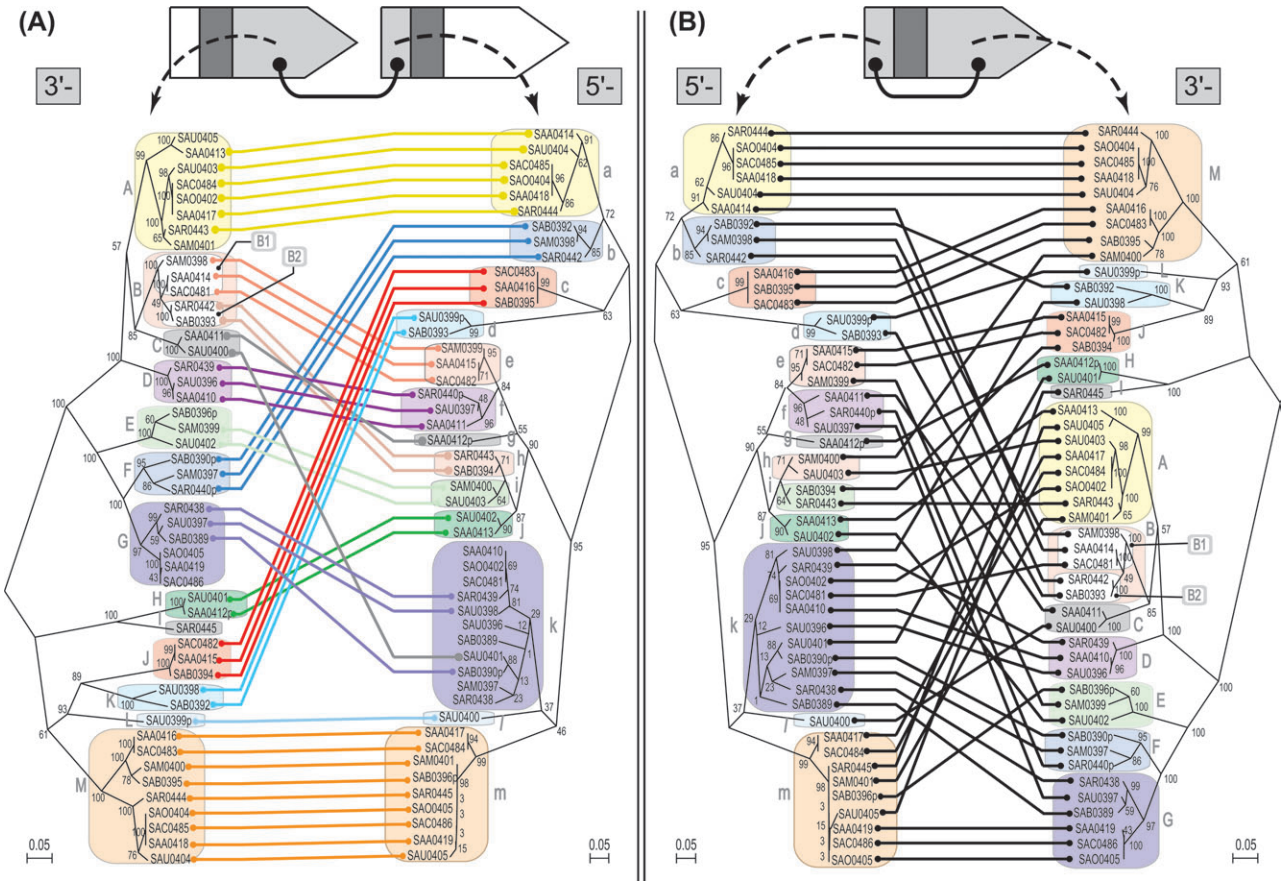

FIG. 4.—Phylogenetic comparison of the 5′ regions and 3′ regions of *lpl* ORFs. A nucleotide NJ phylogeny for 5′-variable regions and one for 3′-variable regions were compared with each other by connecting OTUs in a pair of 3′-variable region of an ORF and 5′-variable region of its downstream ORF (*A*) and in a pair within an ORF (*B*). The bootstrap values (%) were obtained from 1,000 resamplings. Groups of the conserved pairs are indicated in different colors of connecting lines in (*A*). A paraphyletic group, B1, and the other monophyletic groups in each phylogeny corresponding to the groups of the pairs in (*A*) are indicated by boxes both in (*A*) and (*B*).

has been maintained, their intervening sequence should also have been conserved (fig. 3). To verify this, the intergenic regions were compared by multiple alignments (fig. 5). They turned out to be highly conserved within each group (see also table 1, intergenic region). Here "A–a" represents, for example, an intergenic region sandwiched by 3′-variable region of "A" and 5′-variable region of "a".

In most of the cases, the sequences from different groups differed in length and composition, showing almost no similarity. A putative ribosome-binding site (Novick 1991) could be identified in all their sequences except for "K–d," and their sequences from the different groups are likewise distinctive. There are, however, several cases in which the intergenic sequences from different groups are similar to each other: among "B1–e," "B2–h," and "D–f"; between "F–b" and "G–k"; and between "H–j" and "L–l" (fig. 5). Additionally, "C–k" and "C–g," the exceptional pairs of the conservation, also showed similarity to each other and to "A–a".

Therefore, the conserved pair of the two variable regions can be extended to the conserved unit comprising a 3′-variable region, a downstream intergenic region, and a downstream 5′-variable region, which spans two adjacent ORFs. The units from different groups are substantially distinctive from each other, though some sequence families of intergenic regions are found to be common to a few different groups.

Intergenomic Comparison and Reconstruction of Genome Rearrangement Events

The result of the grouping based on the phylogeny comparison in figure 4*A* is displayed in a schematic map of ORFs in figure 6*A*. Gene orders and gene compositions are highly variable among strains, which indicates the occurrence of multiple rearrangement events in this region in the past. Combinations of the 5′-variable region and the 3′-variable region within an ORF appear to have been extensively shuffled. For instance, "f" is paired with "G" in SAU0397/SAV0434 of N315/Mu50, but it is paired with "C" in SAA0411 of USA300. At the level of ORF, 32 distinct patterns were generated in the combination of each 13 groups of 3′ and 5′ regions (fig. 6*A*). In contrast, the presence of the conserved pairs of the 3′-variable region of one ORF and the 5′-variable region of its downstream ORF, as displayed by the same coloring, can be clearly observed here. These observations also support the hypothesis that the linkage encompassing two adjacent genes has been well maintained in spite of the extensive rearrangements.

Occurrence of numerous rearrangements of various types can be inferred from pairwise genome comparisons. For example, 1) an indel is found between USA300 (k-D-f-C-g-H-j-A-a-B1-) and COL (k-B1-), with an apparent deletion of "D-f-C-g-H-j-A-a." 2) A translocation is found between N315/Mu50 (-D-f-G-k-) and MSSA252 (-G-k-D-f-). 3) A substitution is found between RF122 (-b-K-d-B2-h-J-c-M-) and MW2/MSSA476 (-b-B1-e-E-i-M-). In addition, 4) the presence of two units of A–a in USA300 indicates an apparent gene conversion (or a replicative transposition).

Among these rearrangements, two indels, one found between USA300 and COL and the other between COL and NCTC8325, are remarkable in that the variation between the strains can be explained by one deletion event. The close relationship between these three strains has been elucidated by phylogenetic analysis using the sequences of housekeeping genes (Baba et al. 2008).

For the above two indels, alignment of the sequences involved allowed determination of the recombination site (fig. 6*B* and *C*). In the upstream region, a progeny sequence in the second line of the three perfectly aligned with one of the parental sequences in the first line but not with the other parental sequence in the third line. Then in the downstream region, the progeny sequence in the second line aligns perfectly with the parental sequence in the third line but not with the parental sequence in the first line. The transient region, around which recombination is likely to have taken place, coincides with the central conserved regions in both the cases. This indicates that these indels have been generated by a homologous recombination event at the central conserved region.

A Model of Paralog Cluster Diversification through Homologous Recombination between Central Conserved Regions

Using the observations and arguments presented above, we propose a model for the diversification processes in this tandem paralog cluster (fig. 3 and fig. 7). During the sequence diversification process, the central conserved region somehow maintained its sequence being sandwiched by two variable regions. The central conserved region has provided a site for mutual homologous recombination. A recombination event was able to change combination of the 5′-variable region and the 3′-variable region within one ORF but unable to disturb the linkage of the 3′-variable region of one ORF, the intergenic region, and the 5′-variable region of its downstream ORF, resulting in maintenance of this conserved unit (fig. 3). These processes make it possible to generate an ORF with a novel combination of the two variable regions.

Multiple rounds of such crossing-over can cause various types of rearrangements including deletions, conversions, translocations, and substitutions, which are observed among the genomes studied (fig. 6*A*). The occurrence of translocation can be explained by circle formation and ensuring reintegration, which is proposed as a mechanism to cause gene amplification or phase variation (Mahan and Roth 1989; Howell-Adams and Seifert 2000; Barten and Meyer 2001), though multiple rounds of unequal crossing-over events can also result in translocation. An apparent gene conversion can be explained by repeated rounds of crossing-over (Yamamoto et al. 1988, 1992).

If horizontal gene transfer takes place between closely related strains, intermolecular recombination might also occur (Hacker and Kaper 2000; Ochman et al. 2000) and could contribute to these rearrangements, as illustrated in figure 7*B*. Genetic exchange via horizontal gene transfer has been suggested by sequence comparison of the *hsdS* gene of the Type I restriction–modification

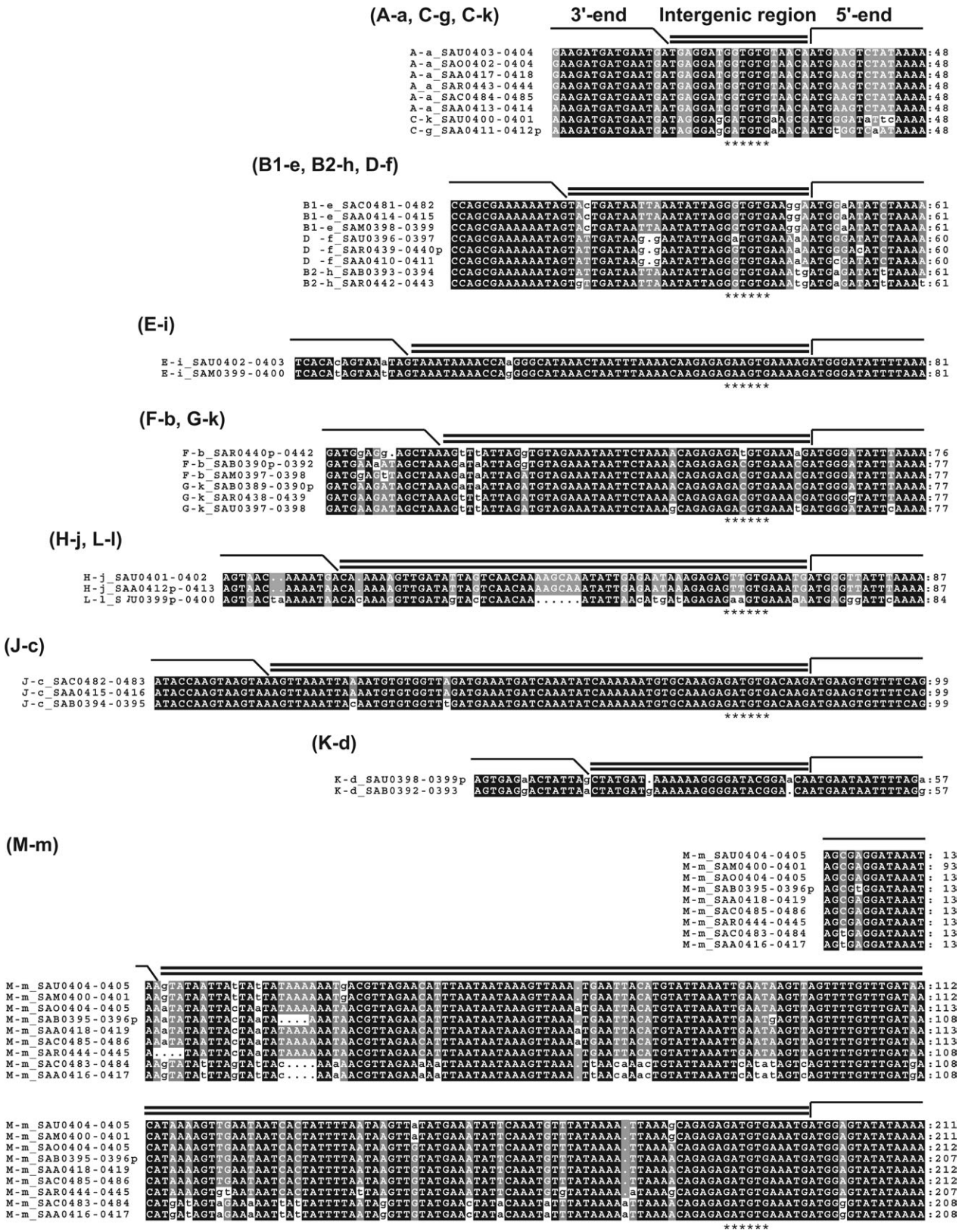

FIG. 5.—Multiple alignments of the *lpl* intergenic regions. "A–a" represents, for example, an intergenic region sandwiched by the 3′-variable region of "A" group and the 5′-variable region of "a" group. A putative ribosome-binding site (Novick 1991), which could be found in all except for "K–d", is indicated by asterisks.

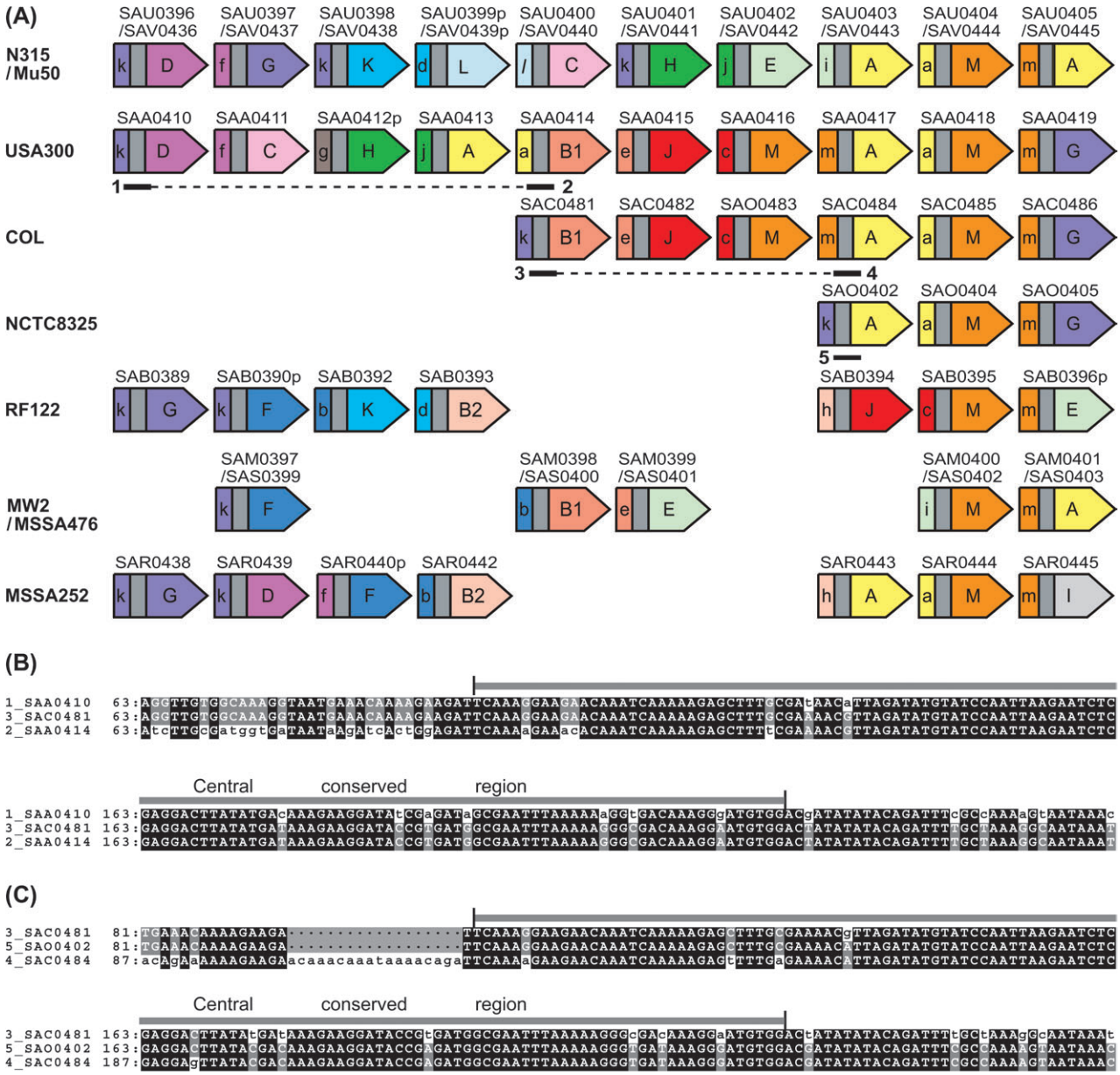

FIG. 6.—(*A*) Schematic maps of *lpl* ORFs. Naming and coloring for 5′-variable region and 3′-variable region are after the grouping in figure 4*A*. For an indel found between USA300 and COL and one between COL and NCTC8325, an apparently deleted region and the regions involved in recombination are indicated by dotted lines and numbered thick lines, respectively. (*B*) Alignment of the regions 1, 2, and 3 in (*A*) suggesting a recombination relationship between them ($1 \times 2 \rightarrow 3$). The central conserved region is indicated above the alignment. (*C*) Alignment of the regions 3, 4, and 5 in (*A*) suggesting a recombination relationship between them ($3 \times 4 \rightarrow 5$).

system, which is linked to the *lpl* cluster (Tsuru et al. 2006). This genomic island, however, has been considered to be no longer mobile because it harbors only the remnants of an integrase homologue (Baba et al. 2004). Another type of genomic island of *S. aureus*, called SaPIs, has intercellular mobility with the aid of a specific helper phage (Ruzin et al. 2001; Novick and Subedi 2007), but the helper phage for the νSaα island has not yet been identified. Natural transformation has not been reported in *S. aureus*. Occurrence of conjugation was proposed in order to explain large-scale chromosome replacement, suggested from multilocus sequence typing (Robinson and Enright 2004).

It is likely that the diversity of this region was generated through accumulation of these intragenomic and intergenomic processes. This model can explain both formation of highly distorted gene orders and shuffling of the two variable regions of an ORF.

Search for Other Tandem Paralogous Gene Clusters in *S. aureus* with the Same Diversification Pattern as the *lpl* Cluster

One might expect that the present model for diversification is general to paralog clusters: any highly conserved

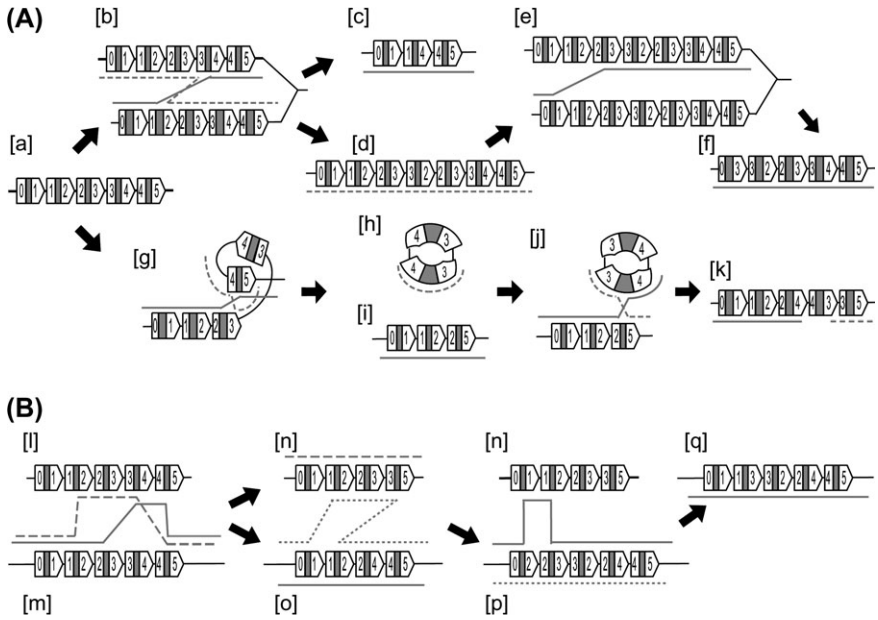

FIG. 7.—Formation of various types of rearrangements is explained by multiple rounds of crossing-over events at the central conserved region (gray bar). (*A*) An unequal crossing-over between sister chromosomes ([b]) will cause deletion ([a] to [c]) and duplication ([a] to [d]). An additional unequal crossing-over ([e]) will result in apparent conversion ([a] to [f]). An intra-chromosomal, unequal crossing-over ([g]) can form a circle ([h]) and a deletion ([a] to [i]), and ensuing re-integration of the circle ([j]) will results in apparent translocation ([a] to [k]). The two routes of deletion formation ([a] to [c]; [a] to [i]) can result in apparent substitution ([c] and [i]). (*B*) Inter-molecular recombination involving horizontal gene transfer also explains formation of various types of rearrangements. Double cross-overs between a donor [l] and a recipient [m] will cause deletion ([m] to [n]; [m] to [o]), resulting in apparent substitution ([n] and [o]). Additional inter-molecular recombination between [n] and [o] can result in an apparent conversion ([m] to [p]). Another round of recombination between [p] and [n] can result in an apparent translocation ([m] to [q]).

sequence within tandem paralogous genes would create a site for recombination and could cause similar diversification to the *lpl* cluster. In order to explore this possibility, we searched signs of this kind of rearrangement in other tandem gene clusters.

Two other tandem clusters on the genomic islands of *S. aureus*, the *ssl* cluster on νSaα and the *spl* cluster on νSaβ, were examined in our previous study (Tsuru et al. 2006), which revealed that the possible recombination sites are located at the upstream region through the 5′-terminal region of the genes in both cases. The mechanism and evolutionary consequence of their diversification are thus different from those for the *lpl* genes.

We searched for other tandem paralogous genes in strain N315 of *S. aureus* by performing a self–self genome comparison in CGAT (Uchiyama et al. 2006), which is based on all-against-all BlastN analysis and a local program to detect linked repetitive sequences. This screening uncovered 15 tandem clusters (table 2), which include the *ssl*, *lpl*, and *spl* clusters. We then carried out multiple sequence alignment within member genes of each cluster to examine whether they carry a highly conserved sequence sandwiched by variable sequences. One tandem cluster, composed of SAU1317, SAU1318, SAU1319, and SAU1321 in N315 (encoding hypothetical proteins) showed such a structure.

We carried out screening of their homologs by BlastN search (*e* value cutoff of 10^{-5}) against prokaryote complete genome sequences in the NCBI Genome database (31 July 2007 data). This screening revealed multiple (two or four) genes in all the nine *S. aureus* strains analyzed and a single homologous gene in each of the sequenced *S. epidermidis* strains and in the *S. haemolyticus* strain. Some of these genes were annotated as lipoproteins (Holden et al. 2004), although not all the genes encode a cysteine residue around the signal peptide-like sequence, a hallmark of bacterial lipoproteins (Sibbald et al. 2006) (data not shown). These genes are termed SA1317 homologs in this study. The multiple homologous genes in the *S. aureus* strains are present in tandem, sometimes with intervening ORFs and/or an intervening prophage. Similarity plots using all the sequences of the *S. aureus* genomes confirmed the presence of the central conserved region that is sandwiched by two variable regions (fig. 8).

A phylogenetic comparison was carried out to detect a linkage encompassing two adjacent paralogous genes (fig. 9). The tree for the 5′-variable region is not as well resolved as that for the 3′-variable region, presumably because of its short and variable length. The two phylogenetic trees seem to be more congruent for pairs encompassing adjacent genes (fig. 9*A*) than for pairs within the same gene (fig. 9*B*). We could not apply clear-cut grouping based on the congruence of the two trees for this cluster. Therefore, we grouped 5′-variable region and 3′-variable region separately based on each tree. The 3′-variable region was grouped into seven groups named "T" through "Z," whereas the six 5′-variable region groups were named "u" through "z," such that the evolutionary distance remains equal to or shorter than 0.15 within a group (fig. 9). We then tried to connect these 3′ groups and 5′ groups. This grouping is also indicated in the schematic map in figure 10. This revealed the presence of sequence

Table 2
Tandem Gene Clusters in *Staphylococcus aureus* N315 Genome

Cluster Name	Genes in N315
Lipoprotein-like (Lpl; Locus I)	SAU0092, SAU0092, SAU0093, SAU0094, SAU0095, SAU0096
Hypothetical protein	SAU0282, SAU0286, SAU0287, SAU0288, SAU0289, SAU0290
Superantigen-like (Ssl; νSaα)	SAU0382, SAU0383, SAU0384, SAU0385, SAU0386, SAU0387, SAU0388, SAU0389, SAU0390
Lipoprotein-like (Lpl; νSaα)	SAU0396, SAU0397, SAU0398, SAU0399, SAU0400, SAU0401, SAU0402, SAU0403, SAU0404, SAU0405
Ser–Asp–rich fibrinogen-binding protein	SAU0519, SAU0520, SAU0521
Superantigen-like protein	SAU1009, SAU1010, SAU1011
ECM-binding protein homologue	SAU1267, SAU1268
Hypothetical protein	SAU1317, SAU1318, SAU1319, SAU1321
Serine protease-like (Spl; νSaβ)	SAU1627, SAU1628, SAU1629, SAU1630, SAU1631
Enterotoxin	SAU1642, SAU1643, SAU1644, SAU1645, SAU1646, SAU1647, SAU1648
Hemolysin	SAU2207, SAU2208, SAU2209
Hypothetical protein	SAU2263, SAU2264, SAU2265
Lipoprotein-like (Lpl; Locus III)	SAU2269, SAU2273, SAU2274, SAU2275
Fibronectin-binding protein	SAU2290, SAU2291

conservation in the pairs of two variable regions (figs. 9 and 10). Fisher's exact test supported the significance of the linkage between these pairs; the *P* value is 0.000104.

Intergenomic comparison showed that the sequences of this cluster are similar to each other between N315 and Mu50; between USA300, COL, and NCTC8325; and between MW2 and MRSA476 (fig. 10). This relatedness among the strains is also indicated in the phylogenetic trees shown in figure 9. Comparison between USA300, COL, and NCTC8325 showed that the linkage between "W" and "u" has been conserved regardless of the apparent insertion of prophages in USA300 and NCTC8325. The presence of intervening nonhomologous ORFs and/or an intervening prophage between SA1317 homologs does not seem to distort the linkage spanning two paralogs between the above closely related strains. On the other hand, comparison between the diverged strains did not provide direct evidence of their rearrangements.

Taken together, a significant relationship between the two variable regions encompassing genes was also identified also in the SA1317 homolog tandem cluster. However, we could not judge whether this linkage is formed by the diversification processes proposed in this study (figs. 3 and 7) because the current data sets provided no evidence of rearrangements.

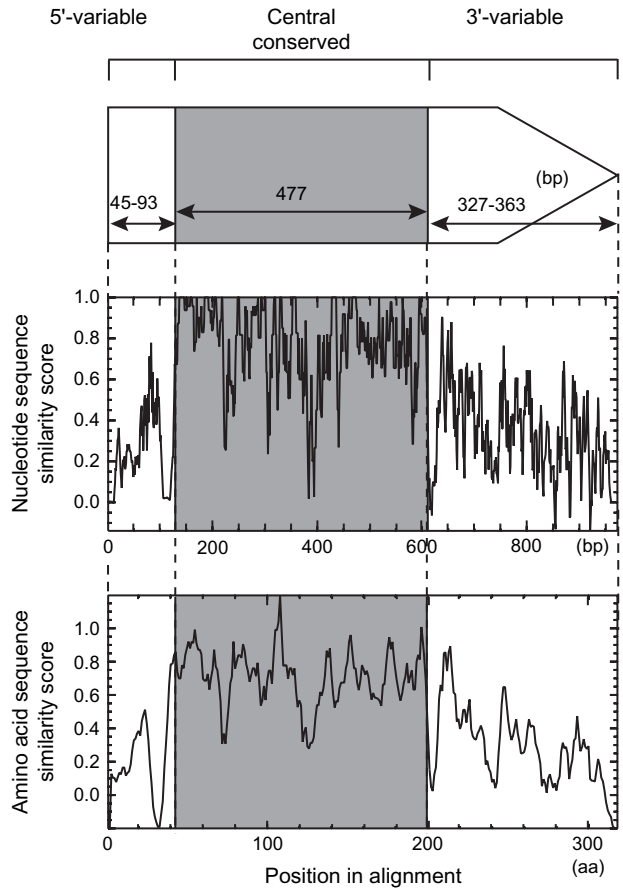

FIG. 8.—Structure of genes of SA1317 homologs (top) with similarity plots for nucleotide sequences (middle) and for amino acid sequences (bottom). A central conserved region is highlighted by gray shading.

Discussion

The phylogenetic trees for the 3′-variable region of the *lpl* ORF and the 5′-variable region of its downstream ORF (fig. 4*A*) appeared congruent, but the congruence is not complete. This may be because the level of divergence among the groups is too extensive to reconstruct their evolutionary relationships in either or both of the trees (Nei and Kumar 2000). Another (not mutually exclusive) possibility is that the 5′-variable region may be too short to construct a reliable tree (Nei and Kumar 2000).

The conserved units comprising the 3′-variable region, the intergenic region, and the 5′-variable region are maintained within a group in almost all the cases (figs. 4*A* and 6). There are, however, some exceptions in their one-to-one correspondence. One example is present in the 3′-variable region of "B"; there are two types, "B1" and "B2", which are paired with "e" and "h" groups of the 5′-variable region, respectively. A similar situation is found in the 5′-variable region of "k". "k" is paired with "G" group of the 3′-variable region in most

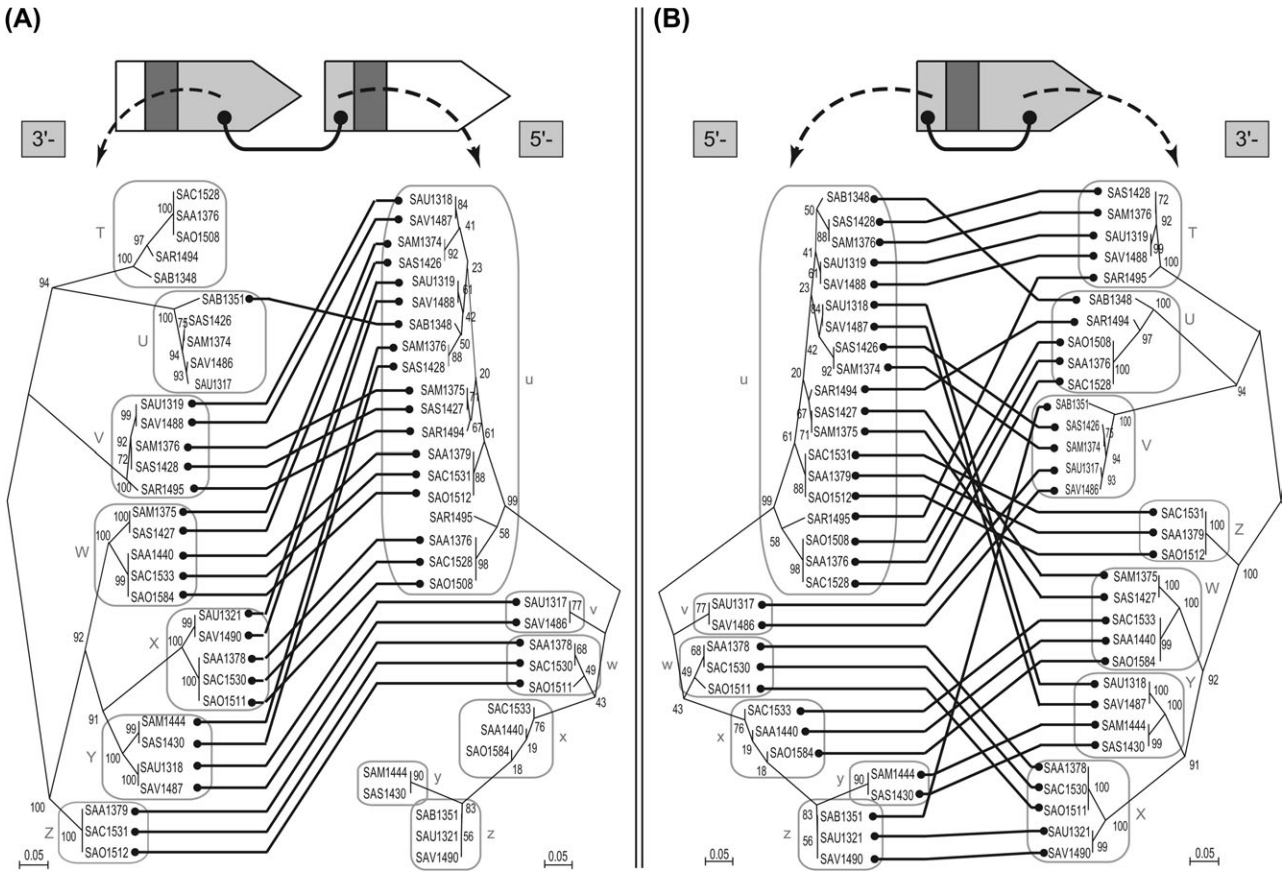

FIG. 9.—Phylogenetic comparison of the 5′ regions and 3′ regions of SA1317 homologs. A nucleotide NJ phylogeny for 5′-variable regions and one for 3′-variable regions were compared with each other by connecting OTUs in a pair of 3′-variable region of an ORF and 5′-variable region of its downstream ORF (*A*) and in a pair within an ORF (*B*). The bootstrap values (%) were obtained from 1,000 resamplings. Groups of each phylogeny were assigned so as the mutual evolutionary distances remain equal to or shorter than 0.15 within a group.

cases. However, in SAU0400/SAV0440–SAU0401/SAV0441 of N315/Mu50, "k" is paired with "C". Such imperfect conservation of the unit suggests the involvement of mechanisms other than the homologous recombination at the central conserved region. Illegitimate recombination that requires only short or no homology may have taken place at their intergenic region. The putative recombination points lie close to the 5′ end in both the cases.

The 3′-variable region of "C" group is involved in another exceptional pair: "C" is paired with "k" in SAU0400/SAV0440–SAU0401/SAV0441 of N315/Mu50 as mentioned above but paired with "g" in SAA0411–SAA0412p in USA300 (fig. 4*A*, gray connecting lines; fig. 6*A*). The prototype of the 5′-variable region of SAA0412p, a pseudogene, could be "k". Fast accumulation of mutations in the 5′-variable region of the ancestor of SAA0412p may have led to its erroneous grouping into "g". Sequence similarity in their intergenic regions (fig. 5) and the common occurrence of "H–j" pair in their downstream (fig. 6) support this hypothesis.

One prediction from our diversification model is that the 5′-variable region at the upstream end of this cluster and the 3′-variable region at the downstream end should be conserved among strains. In the case of the *lpl* cluster, conservation of "k" at the upstream end is consistent with this prediction (fig. 6*A*). However, the 3′-variable region at the downstream end is not conserved (fig. 6*A*), which indicates the involvement of a mechanism other than our model. Illegitimate recombination could be involved in the diversification of the downstream end. Interestingly, such a recombination event that involves a long homology at one joint and a short homology in the other joint has been reported in various bacteria (Kusano et al. 1997; de Vries and Wackernagel 2002; Prudhomme et al. 2002).

The present model attempts to explain the generation of a novel paralog by changing the combination of the two variable regions. It does not address the issue of the origin of the sequence diversity among these groups. The intergroup sequence diversity contrasts with the intragroup sequence conservation. Thus, it seems reasonable to assume that these represent two distinctive processes. The genomic island carrying this *lpl* cluster is likely to have been acquired by an ancestor of *S. aureus* because this island is common to this species but has not been found in any other *Staphylococcus* species (Baba et al. 2004). The diverse repertoire of the *lpl* genes may have been formed before the island was acquired; yet, we do not know the precise molecular mechanisms generating the diverse repertoire.

Bacterial surface proteins are often more variable in their amino acid sequence than proteins encoded by housekeeping genes, presumably due to diversifying selection, exerted by the host immune system (Caporale 2003). In

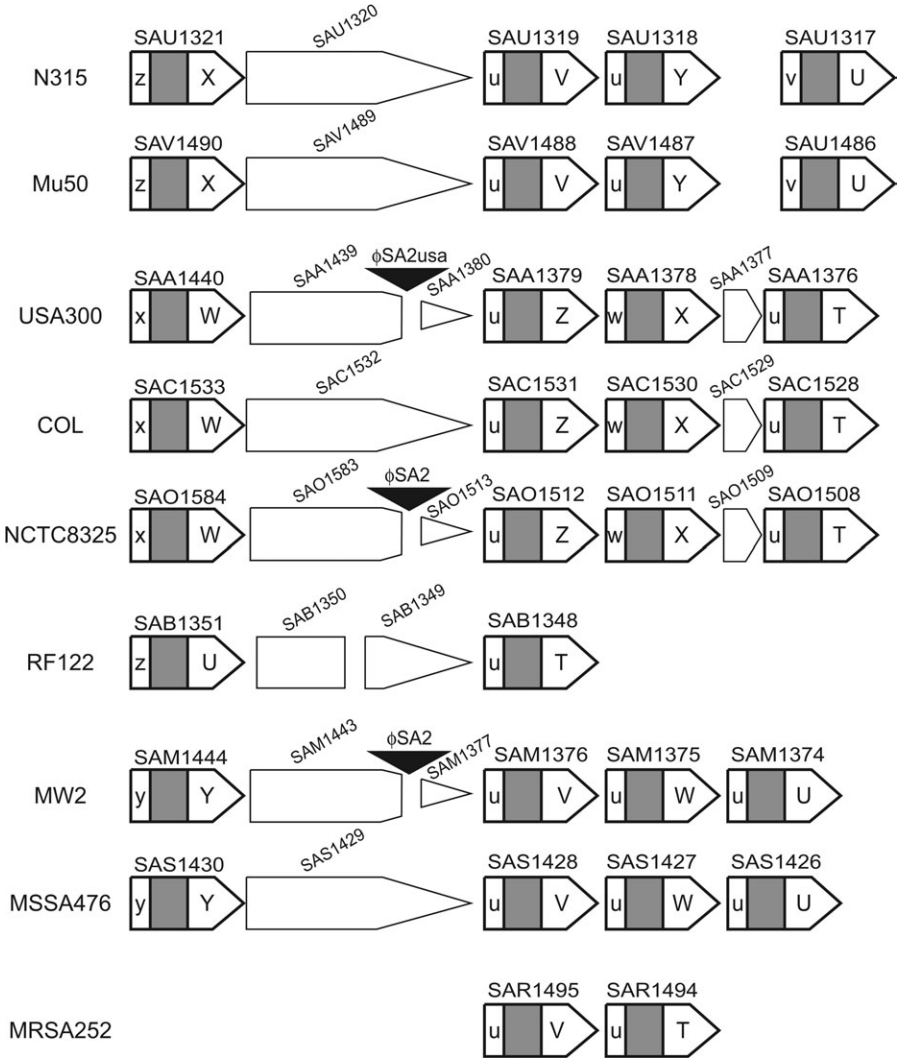

FIG. 10.—Maps of SA1317 homolog clusters in various *Staphylococcus aureus* strains. The SA1317 homologs are drawn in bold lines. Naming of their 5′-variable region and 3′-variable region is after the tree-based grouping in figure 9. The larger intervening ORF, SAU1320 and its homologous genes, is observed in all the strains except for MRSA252, whereas the shorter intervening ORF, SAA1377 and its homologs, is observed in strains USA300, COL, and NCTC8325. SAB1350 and SAB1349 in RF122 are truncated genes homologous to SAU1320. Insertion of a prophage into the larger intervening ORF observed in USA300, NCTC8325, and MW2 is indicated by a black triangle.

S. aureus, IgG-binding protein A gene (*spa*) and seven *S. aureus* surface protein genes (*sas* genes) are known to be so and have been used for strain typing (Shopsin et al. 1999; Mazmanian et al. 2001). In order to examine the evolutionary rate of the *lpl* genes, we chose SAU0404 (in N315), SAA0418 (in USA300), and SAR0444 (in MRSA252), as a possible orthologous gene set in the *lpl* cluster (figs. 4 and 6), and compared their evolutionary distances with those of the seven concatenated housekeeping genes (*arc*, *aroE*, *glpF*, *gmk*, *pta*, *tpi*, and *yqiL*) used in multilocus sequence typing (Enright et al. 2000). The distances calculated for the *lpl* genes are much longer than those calculated for the concatenated sequences of the seven housekeeping genes: the average Kimura's two-parameter distance (Kimura 1980) for the *lpl* genes is 0.066 compared with 0.0077 for the housekeeping genes. Phylogenies based on these distances are presented in supplementary figure S1*B* (Supplementary Material online). This result indicates that the *lpl* genes have a relatively fast diversification rate. This fast sequence diversification, together with shuffling via recombination as proposed in our model, has very likely contributed to variability in this cluster.

The sequence divergence/conservation was detected using the whole sequences of the *lpl* region in this study. One might expect that a small portion might undergo gene conversion between different groups. To further explore this possibility, we performed detailed pairwise nucleotide sequence comparisons using Blast2 (Tatusova and Madden 1999). These results detected several examples indicating gene conversion (Tsuru T, Kobayashi I, unpublished data). However, contribution of such gene conversion seems so small that it did not affect the clear grouping in figure 4*A*. Likewise, the presence of significantly related *lpl* homologs in the loci of *S. aureus* other than νSaα and in some other *Staphylococcus* species leaves the possibility of interloci recombination and interspecies recombination. The

same detailed Blast2 analysis, however, did not detect any sequence homology indicating such interactions; the maximal length of the sequence homology equal or more than 95% identity was only 48 bp. Therefore, it is likely that the *lpl* genes of *S. aureus* have evolved separately at each locus, as was suggested by phylogeny analysis using full-length sequences in supplementary figure S1 (Supplementary Material online).

The sequences of the central conserved region defined by us are not always completely identical to each other along the entire length (table 1 and supplementary fig. S3 [Supplementary Material online]). This suggests that the precise recombination point varies in each event within the central conserved region because the different recombination points will explain the generation of sequence diversity of the central conserved region. In this study, we were able to estimate the recombination site for the two indels found among USA300, COL, and NCTC8325 (fig. 6). In both these cases, the detected recombination points resided within the central conserved region, but their positions were different from each other. If we examined more sequences and could find recombination events between them, we might be able to test the above hypothesis. In addition, comparison of the sequence of very closely related strains will enable us to determine the position of recombination more precisely. Recently, entire genome sequences have become available for 4 more *S. aureus* strains: JH1 and JH9 (Mwangi et al. 2007), USA300-TCH156 (Highlander et al. 2007), and Newman (Baba et al. 2008). Additionally, ten genomes of USA300 derivatives were examined by comparative genome sequence analysis (Kennedy et al. 2008). These are closely related to each other or to the other sequenced strains. Although we studied the sequences of these strains to obtain more examples of recombination events, we could not obtain any additional insight into the *lpl* gene cluster: JH1 and JH9 are almost identical to N315/Mu50 with only a few point mutations throughout the relevant region, and USA300-TCH156 and Newman are almost identical to USA300. From this analysis, we concluded that testing the above hypothesis would require many more genome sequences of appropriate divergence.

The frequent recombination at the central conserved region of the *lpl* genes on νSaα suggests, on the other hand, that an intrastrain homogenization process might occur in this region. However, we could not detect such tendency in the phylogenetic analysis using the sequences of the central conserved region (data not shown).

Why has the central region been conserved? One possibility is that this region has been maintained because of its functional importance. At present, neither a functional role for the central conserved region nor a physiological role of these *lpl* paralogs has been discovered. Generally, bacterial lipoproteins are involved in cellular processes such as antibiotic resistance, adhesion to host cells or other bacterial cells, transport of substances, or intercellular communication (Sibbald et al. 2006). Several lipoproteins are recognized as an antigen by a host's immune system, and they thus affect bacterial survival and their pathogenicity (Henderson et al. 1996). Antigenic variation systems involving lipoproteins have been identified and characterized in *Borrelia* (Haake 2000; Norris 2006), *Porphyromonas* (Hall et al. 2005), and *Mycoplasma* (Lysnyansky et al. 2001; Glew et al. 2002; Ron et al. 2002). Tandemly encoded lipoproteins such as *vsp* lipoproteins in *Borrelia hermsii* (Restrepo et al. 1994; Barbour and Restrepo 2000) or *p35* family lipoproteins in *Mycoplasma penetrans* (Neyrolles, Chambaud, et al. 1999; Neyrolles, Eliane, et al. 1999) cause antigenic variation. We do not know whether the interstrain variation in the *lpl* gene arrangement represents a type of antigenic variation, but we think it is likely because this explains why many copies of these paralogs have been maintained (Jordan et al. 2001; Hooper and Berg 2003). The relatively high evolutionary rate observed in this cluster also supports this view. If this is the case, the central region may be conserved for its function as a recombination site for programed diversification similarly to DNA repeats in some other antigenic variation systems. Shuffling of two variable regions may contribute to alteration of sets of Lpl proteins on the cell surface, partly because replacement of the signal peptides in the 5′-variable region will affect efficiency of secretion of each lipoprotein (Sibbald et al. 2006). The wide diversity in the sequence of the intergenic region including a putative ribosome-binding site (fig. 5) might contribute to diversity in expression efficiency of these genes. Recent transcriptome analyses using the sequenced strains have reported expression of some of their *lpl* genes (http://www.bioinformatics.org/sammd/; Sobral et al. 2007), although this has not been confirmed by proteome analyses (Nandakumar et al. 2005; Gatlin et al. 2006). How their expression is regulated has not been elucidated. Precise analysis to identify their differential expression within a strain and between strains is necessary to understand their biological role.

A prominent feature of our model of the diversification process is the presence of the conserved unit encompassing adjacent ORFs. Mosaic gene formation by intergenomic and intragenomic recombination has been reported in various bacteria for antibiotic resistance genes (Maiden 1998; Normark BH and Normark S 2002) and antigen genes (Bessen and Hollingshead 1994; Hobbs et al. 1998; Lachenauer et al. 2000). For allelic variation in the *por* gene for porin protein in *Neisseria gonorrhoeae*, homologous recombination of a partial conserved sequence within the gene is proposed to be responsible for diversity (Cooke et al. 1998; Fudyk et al. 1999; Hamilton and Dillard 2006). However, a linkage of segments beyond adjacent ORFs has not been analyzed so far. The present characteristic mode of tandem paralog diversification, maintaining the 3′ part of a gene, the intergenic region, and the 5′ part of its downstream gene as a unit of evolution, is, to our knowledge, novel among studies of paralog rearrangements.

How general is the presented model of diversification among tandemly repeated genes? A further search for other tandem paralog clusters on *S. aureus* genome revealed another tandem gene cluster, SA1317 homolog, with a gene structure in which a highly conserved sequence is sandwiched between variable sequences (fig. 8). A significant linkage between the variable regions encompassing the genes was also identified in this cluster (figs. 9 and 10), although whether this linkage is formed by the present diversification mechanism is not clear because no evidence of rearrangement was found there. We also examined the other

well-characterized tandem paralog genes in other bacteria. In the *p35* lipoprotein genes in *M. penetrans*, a conserved region is located at their 5′ end encoding a signal peptide (Sasaki et al. 2002). In the *vsp* genes in *Mycoplasma bovis*, a conserved sequence is found upstream of the ORF (Lysnyansky et al. 2001). In both these cases, a different diversification process from our model must be operating.

Conclusion

We tend to regard an ORF as a unit of gene evolution as well as a unit of gene expression and its function. The present work with a tandem paralog cluster identified a unit consisting of the 3′ half of a gene, a downstream intergenic region, and the 5′ half of a downstream gene as the unit of evolution. This is because the central region of the gene provides a site to recombine the 5′ half and the 3′ half of a gene to generate variation.

Supplementary Material

Supplementary table S1, figures S1–S3, and legends for the supplementary figures are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Dr Tadashi Baba and Dr Ikuo Uchiyama for discussion of the work. This work was supported by the 21st century Center of Excellence projects of "Elucidation of Language Structure and Semantic behind Genome and Life System" from Ministry of Education, Culture, Sports, Science, and Technology to I.K.; "Grants-in-Aid for Scientific Research" from Japan Society for the Promotion of Science (JSPS) to I.K. (19790316, 19657002) and T.T. (19-10124); The NIBB (National Institute for Basic Biology) Cooperative Research Program (8-203); and Cosmetology Research Grants from The Cosmetology Research Foundation to I.K. T.T. is a JSPS Research Fellow (DC2).

Literature Cited

Agresti A. 1992. A survey of exact inference for contingency tables. Stat Sci. 7:173–177.

Baba T, Bae T, Schneewind O, Takeuchi F, Hiramatsu K. 2008. Genome sequence of *Staphylococcus aureus* strain Newman and comparative analysis of staphylococcal genomes: polymorphism and evolution of two major pathogenicity islands. J Bacteriol. 190:300–310.

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

Baba T, Takeuchi F, Kuroda M, Yuzawa H, Ito T, Hiramatsu K. 2004. The genome of *Staphylococcus aureus*. In: Ala'Aladeen DAA, Hiramatsu K, editors. The *Staphylococcus aureus*: molecular and clinical aspects. London: Eliis Harwood. p. 66–153.

Barbour AG, Restrepo BI. 2000. Antigenic variation in vector-borne pathogens. Emerg Infect Dis. 6:449–457.

Barten R, Meyer TF. 2001. DNA circle formation in *Neisseria gonorrhoeae*: a possible intermediate in diverse genomic recombination processes. Mol Gen Genet. 264:691–701.

Bessen DE, Hollingshead SK. 1994. Allelic polymorphism of *emm* loci provides evidence for horizontal gene spread in group A streptococci. Proc Natl Acad Sci USA. 91:3280–3284.

Caporale LH. 2003. Natural selection and the emergence of a mutation phenotype: an update of the evolutionary synthesis considering mechanisms that affect genome variation. Annu Rev Microbiol. 57:467–485.

Chavakis T, Preissner KT, Herrmann M. 2007. The anti-inflammatory activities of *Staphylococcus aureus*. Trends Immunol. 28:408–418.

Clarkson DB, Fan YA, Joe H. 1993. A remark on algorithm-643-fexact—an algorithm for performing Fishers exact test in r x c contingency-tables. ACM Trans Math Softw. 19:484–488.

Cooke SJ, Jolley K, Ison CA, Young H, Heckels JE. 1998. Naturally occurring isolates of *Neisseria gonorrhoeae*, which display anomalous serovar properties, express PIA/PIB hybrid porins, deletions in PIB or novel PIA molecules. FEMS Microbiol Lett. 162:75–82.

de Vries J, Wackernagel W. 2002. Integration of foreign DNA during natural transformation of *Acinetobacter* sp. by homology-facilitated illegitimate recombination. Proc Natl Acad Sci USA. 99:2094–2099.

Diep BA, Gill SR, Chang RF, et al. (12 co-authors). 2006. Complete genome sequence of USA300, an epidemic clone of community-acquired meticillin-resistant *Staphylococcus aureus*. Lancet. 367:731–739.

Dobrindt U, Hochhut B, Hentschel U, Hacker J. 2004. Genomic islands in pathogenic and environmental microorganisms. Nat Rev Microbiol. 2:414–424.

Enright MC, Day NP, Davies CE, Peacock SJ, Spratt BG. 2000. Multilocus sequence typing for characterization of methicillin-resistant and methicillin-susceptible clones of *Staphylococcus aureus*. J Clin Microbiol. 38:1008–1015.

Friedman R, Hughes AL. 2001. Gene duplication and the structure of eukaryotic genomes. Genome Res. 11:373–381.

Fudyk TC, Maclean IW, Simonsen JN, Njagi EN, Kimani J, Brunham RC, Plummer FA. 1999. Genetic diversity and mosaicism at the *por* locus of *Neisseria gonorrhoeae*. J Bacteriol. 181:5591–5599.

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

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

Gatlin CL, Pieper R, Huang ST, et al. (12 co-authors). 2006. Proteomic profiling of cell envelope-associated proteins from *Staphylococcus aureus*. Proteomics. 6:1530–1549.

Gevers D, Vandepoele K, Simillon C, Van de Peer Y. 2004. Gene duplication and biased functional retention of paralogs in bacterial genomes. Trends Microbiol. 12:148–154.

Gill SR, Fouts DE, Archer GL, 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.

Glew MD, Marenda M, Rosengarten R, Citti C. 2002. Surface diversity in *Mycoplasma agalactiae* is driven by site-specific DNA inversions within the *vpma* multigene locus. J Bacteriol. 184:5987–5998.

Haake DA. 2000. Spirochaetal lipoproteins and pathogenesis. Microbiology. 146(Pt 7):1491–1504.

Hacker J, Kaper JB. 2000. Pathogenicity islands and the evolution of microbes. Annu Rev Microbiol. 54:641–679.

Hall LM, Fawell SC, Shi X, Faray-Kele MC, Aduse-Opoku J, Whiley RA, Curtis MA. 2005. Sequence diversity and antigenic variation at the *rag* locus of *Porphyromonas gingivalis*. Infect Immun. 73:4253–4262.

Hamilton HL, Dillard JP. 2006. Natural transformation of *Neisseria gonorrhoeae*: from DNA donation to homologous recombination. Mol Microbiol. 59:376–385.

Henderson B, Poole S, Wilson M. 1996. Bacterial modulins: a novel class of virulence factors which cause host tissue pathology by inducing cytokine synthesis. Microbiol Rev. 60:316–341.

Herron-Olson L, Fitzgerald JR, Musser JM, Kapur V. 2007. Molecular correlates of host specialization in *Staphylococcus aureus*. PLoS ONE. 2:e1120.

Highlander SK, Hulten KG, Qin X, et al. (33 co-authors). 2007. Subtle genetic changes enhance virulence of methicillin resistant and sensitive *Staphylococcus aureus*. BMC Microbiol. 7:99.

Hobbs MM, Malorny B, Prasad P, Morelli G, Kusecek B, Heckels JE, Cannon JG, Achtman M. 1998. Recombinational reassortment among *opa* genes from ET-37 complex *Neisseria meningitidis* isolates of diverse geographical origins. Microbiology. 144(Pt 1):157–166.

Holden MT, Feil EJ, Lindsay JA, 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.

Hooper SD, Berg OG. 2003. On the nature of gene innovation: duplication patterns in microbial genomes. Mol Biol Evol. 20:945–954.

Howell-Adams B, Seifert HS. 2000. Molecular models accounting for the gene conversion reactions mediating gonococcal pilin antigenic variation. Mol Microbiol. 37:1146–1158.

Hughes D, Norstrom T. 2005. Biological consequences for bacteria of homologous recombination. In: Mullany P, editor. The dynamic bacterial genome. New York: Cambridge University Press. p. 351–384.

Jordan IK, Makarova KS, Spouge JL, Wolf YI, Koonin EV. 2001. Lineage-specific gene expansions in bacterial and archaeal genomes. Genome Res. 11:555–565.

Kennedy AD, Otto M, Braughton KR, et al. (13 co-authors). 2008. Epidemic community-associated methicillin-resistant *Staphylococcus aureus*: recent clonal expansion and diversification. Proc Natl Acad Sci USA. 105:1327–1332.

Khasanov FK, Zvingila DJ, Zainullin AA, Prozorov AA, Bashkirov VI. 1992. Homologous recombination between plasmid and chromosomal DNA in *Bacillus subtilis* requires approximately 70 bp of homology. Mol Gen Genet. 234:494–497.

Kihara D, Kanehisa M. 2000. Tandem clusters of membrane proteins in complete genome sequences. Genome Res. 10:731–743.

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol. 16:111–120.

Kulig P, Zabel BA, Dubin G, Allen SJ, Ohyama T, Potempa J, Handel TM, Butcher EC, Cichy J. 2007. *Staphylococcus aureus*-derived staphopain B, a potent cysteine protease activator of plasma chemerin. J Immunol. 178:3713–3720.

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

Kuroda M, Yamashita A, Hirakawa H, et al. (13 co-authors). 2005. Whole genome sequence of *Staphylococcus saprophyticus* reveals the pathogenesis of uncomplicated urinary tract infection. Proc Natl Acad Sci USA. 102:13272–13277.

Kusano K, Sakagami K, Yokochi T, Naito T, Tokinaga Y, Ueda E, Kobayashi I. 1997. A new type of illegitimate recombination is dependent on restriction and homologous interaction. J Bacteriol. 179:5380–5390.

Lachenauer CS, Creti R, Michel JL, Madoff LC. 2000. Mosaicism in the alpha-like protein genes of group B streptococci. Proc Natl Acad Sci USA. 97:9630–9635.

Li W-H. 1997. Molecular evolution. Sunderland (MA): Sinauer Associates, Inc.

Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. Science. 290:1151–1155.

Lysnyansky I, Ron Y, Yogev D. 2001. Juxtaposition of an active promoter to *vsp* genes via site-specific DNA inversions generates antigenic variation in *Mycoplasma bovis*. J Bacteriol. 183:5698–5708.

Mahan MJ, Roth JR. 1989. Role of *recBC* function in formation of chromosomal rearrangements: a two-step model for recombination. Genetics. 121:433–443.

Maiden MC. 1998. Horizontal genetic exchange, evolution, and spread of antibiotic resistance in bacteria. Clin Infect Dis. 27(Suppl 1):S12–S20.

Majewski J, Cohan FM. 1998. The effect of mismatch repair and heteroduplex formation on sexual isolation in *Bacillus*. Genetics. 148:13–18.

Mazmanian SK, Ton-That H, Schneewind O. 2001. Sortase-catalysed anchoring of surface proteins to the cell wall of *Staphylococcus aureus*. Mol Microbiol. 40:1049–1057.

Mehta CR, Patel NR. 1986. FEXACT—a fortran subroutine for Fisher exact test on unordered rxc contingency-tables. ACM Trans Math Softw. 12:154–161.

Mwangi MM, Wu SW, Zhou Y, et al. (11 co-authors). 2007. Tracking the in vivo evolution of multidrug resistance in *Staphylococcus aureus* by whole-genome sequencing. Proc Natl Acad Sci USA. 104:9451–9456.

Nandakumar R, Nandakumar MP, Marten MR, Ross JM. 2005. Proteome analysis of membrane and cell wall associated proteins from *Staphylococcus aureus*. J Proteome Res. 4:250–257.

Nei M, Kumar S. 2000. Molecular evolution and phylogenetics. New York: Oxford University Press.

Neyrolles O, Chambaud I, Ferris S, Prevost MC, Sasaki T, Montagnier L, Blanchard A. 1999. Phase variations of the *Mycoplasma penetrans* main surface lipoprotein increase antigenic diversity. Infect Immun. 67:1569–1578.

Neyrolles O, Eliane JP, Ferris S, da Cunha RA, Prevost MC, Bahraoui E, Blanchard A. 1999. Antigenic characterization and cytolocalization of P35, the major *Mycoplasma penetrans* antigen. Microbiology. 145(Pt 2):343–355.

Normark BH, Normark S. 2002. Evolution and spread of antibiotic resistance. J Intern Med. 252:91–106.

Norris SJ. 2006. Antigenic variation with a twist—the *Borrelia* story. Mol Microbiol. 60:1319–1322.

Novick RP. 1991. Genetic systems in staphylococci. Methods Enzymol. 204:587–636.

Novick RP, Subedi A. 2007. The SaPIs: mobile pathogenicity islands of *Staphylococcus*. Chem Immunol Allergy. 93:42–57.

Ochman H, Lawrence JG, Groisman EA. 2000. Lateral gene transfer and the nature of bacterial innovation. Nature. 405:299–304.

Ohno S. 1970. Evolution of gene duplication. New York: Springer-Verlag.

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

Prince VE, Pickett FB. 2002. Splitting pairs: the diverging fates of duplicated genes. Nat Rev Genet. 3:827–837.

Prudhomme M, Libante V, Claverys JP. 2002. Homologous recombination at the border: insertion-deletions and the trapping of foreign DNA in *Streptococcus pneumoniae*. Proc Natl Acad Sci USA. 99:2100–2105.

Reams AB, Neidle EL. 2004. Selection for gene clustering by tandem duplication. Annu Rev Microbiol. 58:119–142.

Reed SB, Wesson CA, Liou LE, Trumble WR, Schlievert PM, Bohach GA, Bayles KW. 2001. Molecular characterization of a novel *Staphylococcus aureus* serine protease operon. Infect Immun. 69:1521–1527.

Restrepo BI, Carter CJ, Barbour AG. 1994. Activation of a *vmp* pseudogene in *Borrelia hermsii*: an alternate mechanism of antigenic variation during relapsing fever. Mol Microbiol. 13:287–299.

Robinson DA, Enright MC. 2004. Evolution of *Staphylococcus aureus* by large chromosomal replacements. J Bacteriol. 186:1060–1064.

Ron Y, Flitman-Tene R, Dybvig K, Yogev D. 2002. Identification and characterization of a site-specific tyrosine recombinase within the variable loci of *Mycoplasma bovis*, *Mycoplasma pulmonis* and *Mycoplasma agalactiae*. Gene. 292:205–211.

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

Sasaki Y, Ishikawa J, Yamashita A, et al. (11 co-authors). 2002. The complete genomic sequence of *Mycoplasma penetrans*, an intracellular bacterial pathogen in humans. Nucleic Acids Res. 30:5293–5300.

Shen P, Huang HV. 1986. Homologous recombination in *Escherichia coli*: dependence on substrate length and homology. Genetics. 112:441–457.

Shopsin B, Gomez M, Montgomery SO, Smith DH, Waddington M, Dodge DE, Bost DA, Riehman M, Naidich S, Kreiswirth BN. 1999. Evaluation of protein A gene polymorphic region DNA sequencing for typing of *Staphylococcus aureus* strains. J Clin Microbiol. 37:3556–3563.

Sibbald MJ, Ziebandt AK, Engelmann S, et al. (11 co-authors). 2006. Mapping the pathways to staphylococcal pathogenesis by comparative secretomics. Microbiol Mol Biol Rev. 70:755–788.

Sobral RG, Jones AE, Des Etages SG, Dougherty TJ, Peitzsch RM, Gaasterland T, Ludovice AM, de Lencastre H, Tomasz A. 2007. Extensive and genome-wide changes in the transcription profile of *Staphylococcus aureus* induced by modulating the transcription of the cell wall synthesis gene *murF*. J Bacteriol. 189:2376–2391.

Takeuchi F, Watanabe S, Baba T, et al. (14 co-authors). 2005. Whole-genome sequencing of *Staphylococcus haemolyticus* uncovers the extreme plasticity of its genome and the evolution of human-colonizing staphylococcal species. J Bacteriol. 187: 7292–7308.

Tamura K, Dudley J, Nei M, Kumar S. 2007. MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. Mol Biol Evol. 24:1596–1599.

Tatusova TA, Madden TL. 1999. BLAST 2 sequences, a new tool for comparing protein and nucleotide sequences. FEMS Microbiol Lett. 174:247–250.

Taylor JS, Raes J. 2004. Duplication and divergence: the evolution of new genes and old ideas. Annu Rev Genet. 38:615–643.

Tsuru T, Kawai M, Mizutani-Ui Y, Uchiyama I, Kobayashi I. 2006. Evolution of paralogous genes: reconstruction of genome rearrangements through comparison of multiple genomes within *Staphylococcus aureus*. Mol Biol Evol. 23:1269–1285.

Uchiyama I, Higuchi T, Kobayashi I. 2006. CGAT: a comparative genome analysis tool for visualizing alignments in the analysis of complex evolutionary changes between closely related genomes. BMC Bioinformatics. 7:472.

van der Woude MW, Baumler AJ. 2004. Phase and antigenic variation in bacteria. Clin Microbiol Rev. 17:581–611 table of contents.

Villemur R, Deziel E. 2005. Phase variation and antigenic variation. In: Mullany P, editor. The dynamic bacterial genome. New York: Cambridge University Press. p. 277–322.

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

Williams RJ, Ward JM, Henderson B, Poole S, O'Hara BP, Wilson M, Nair SP. 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.

Yamamoto K, Kusano K, Takahashi NK, Yoshikura H, Kobayashi I. 1992. Gene conversion in the *Escherichia coli* RecF pathway: a successive half crossing-over model. Mol Gen Genet. 234:1–13.

Yamamoto K, Yoshikura H, Takahashi N, Kobayashi I. 1988. Apparent gene conversion in an *Escherichia coli rec*$^+$ strain is explained by multiple rounds of reciprocal crossing-over. Mol Gen Genet. 212:393–404.

Zhang YQ, Ren SX, Li HL, et al. (29 co-authors). 2003. Genome-based analysis of virulence genes in a non-biofilm-forming *Staphylococcus epidermidis* strain (ATCC 12228). Mol Microbiol. 49:1577–1593.

Takashi Gojobori, Associate Editor

Accepted August 26, 2008