Lateral Transfer of Genes and Gene Fragments in *Staphylococcus* Extends beyond Mobile Elements‡†

Cheong Xin Chan,1‡ Robert G. Beiko,2 and Mark A. Ragan1*

**ARC Centre of Excellence in Bioinformatics and Institute for Molecular Bioscience, The University of Queensland, Brisbane, Queensland 4072, Australia,**1 and Faculty of Computer Science, Dalhousie University, 6050 University Ave., Halifax, Nova Scotia B3H 1W5, Canada2

Received 18 December 2010/Accepted 5 April 2011

The widespread presence of antibiotic resistance and virulence among *Staphylococcus* isolates has been attributed in part to lateral genetic transfer (LGT), but little is known about the broader extent of LGT within this genus. Here we report the first systematic study of the modularity of genetic transfer among 13 *Staphylococcus* genomes covering four distinct named species. Using a topology-based phylogenetic approach, we found, among 1,354 sets of homologous genes examined, strong evidence of LGT in 368 (27.1%) gene sets, and weaker evidence in another 259 (19.1%). Within-gene and whole-gene transfer contribute almost equally to the topological discordance of these gene sets against a reference phylogeny. Comparing genetic transfer in single-copy and in multicopy gene sets, we observed a higher frequency of LGT in the latter, and a substantial functional bias in cases of whole-gene transfer (little such bias was observed in cases of fragmentary genetic transfer). We found evidence that lateral transfer, particularly of entire genes, impacts not only functions related to antibiotic, drug, and heavy-metal resistance, as well as membrane transport, but also core informational and metabolic functions not associated with mobile elements. Although patterns of sequence similarity support the cohesion of recognized species, LGT within *S. aureus* appears frequently to disrupt clonal complexes. Our results demonstrate that LGT and gene duplication play important parts in functional innovation in staphylococcal genomes.

Staphylococci are nonmotile but invasive Gram-positive bacteria that are associated with various pus-forming diseases in humans and other animals. The most prominent pathogenic species in the genus is *Staphylococcus aureus*, of which various strains colonize the nasal passages and skin in humans, causing illnesses that range from minor skin lesions or infections to life-threatening diseases, e.g., meningitis, sepsis (bacteraemia), and toxic shock syndrome (55, 67, 94). The other species of *Staphylococcus*, although lacking genes that encode virulence factors and toxins, are opportunistic pathogens for immunocompromised patients (2, 74, 82).

One of the major problems in the prognosis of staphylococcal infections is the progressive development of resistance in the *Staphylococcus* species to multiple antibiotics (14), e.g., methicillin (36, 102) and vancomycin (23, 84), which has, as in other pathogenic bacteria, been attributed to the susceptibility of the organisms to genetic transfer (22, 83). Lateral genetic transfer (LGT) occurs when the organisms acquire exogenous genetic material that encodes antibiotic resistance (6, 105), and this material becomes established in the lineage, whether by recombination into the genome or, potentially less stably, on an extrachromosomal genetic element (31, 86). In *Staphylococcus*, transfer of genetic material has been shown to be mediated by phage transduction and conjugative transfer (19, 62, 79).

Although a number of studies have examined the frequency of LGT in prokaryotes using rigorous phylogenetic approaches (8, 63, 64, 85, 108), studies of LGT in *Staphylococcus* (11) have been limited. An early analysis based on sequence similarity searches (60) and more-recent genome comparative studies (44, 66) on *S. aureus* have revealed pathogenicity-related and/or extrachromosomal genes (i.e., genomic elements present externally to the chromosome) that are likely to have been acquired via LGT. These mobile genetic elements (MGEs) (65, 66, 69) include pathogenicity islands, plasmids, transposons, and the staphylococcal cassette genomes (SCCs). Many of these genes are associated with virulence, e.g., “superantigen” genes implicated in toxic shock and food poisoning (80), or with antibiotic resistance, e.g., the SCCmec elements that encode methicillin resistance in *S. aureus* (50, 77), suggesting that genetic transfer is a key contributing factor to the evolution of virulence and antibiotic resistance in these species. Based on allelic profiles derived from seven highly conserved housekeeping genes in various *S. aureus* isolates, recent multilocus sequence typing (MLST) studies (28, 31), while grouping diverse isolates into multiple clonal complexes (CCs; http://saureus.mlst.net/), have also shown that chromosomal genes of *S. aureus* are inherited largely vertically, i.e., parent to offspring within a lineage. However, little is known about the extent of gene sharing within and between these CCs of *S. aureus*.

Although recombination has been extensively studied in bacteria (30, 73), large-scale phylogenetic (phylogenomic) studies are typically based on the assumption that the units of genetic transfer are necessarily whole genes. It has been shown, how-
ever, that genetic material integrated via LGT can constitute an entire gene (41, 104), a partial (fragmentary) gene (10, 49, 75), or multiple (entire or fragmentary) adjacent genes, including multigene operons (34, 48, 81). We recently examined the transfer of genes and gene fragments among 144 prokaryote genomes (15, 18), demonstrating that studies focusing entirely on whole-gene transfer are likely to underestimate the extent of LGT. In the previous analyses, we included only single-copy gene sets to avoid complications of paralogy in the inference of LGT. Using a more-focused data set than in our previous studies, here we apply a statistically based phylogenetic approach to examine the modularity of LGT in 13 completely sequenced genomes of Staphylococcus (both chromosomes and extrachromosomal elements; Table 1), this time also including gene sets with two or more members in individual genomes.

The genomes of S. aureus subsp. aureus N315 and Mu50 were the first of any Staphylococcus spp. to be sequenced and released (60); both strains were isolated from male patients with postsurgical wound infections in Japan. The strains MW2 (4) and MSSA476 (45) are both isolates of community-acquired S. aureus infections; the strain COL (36) is the oldest isolate of methicillin-resistant S. aureus (MRSA) and dates back to 1976. The strain USA300 (24) was isolated from unassociated outbreaks of S. aureus infections in the United States, Canada, and Europe within the last decade. Finally, the strain NCTC8325 (47) has been used as the generic representative strain of S. aureus in most genetic studies. The S. aureus strain MRSA252 (45) is a hospital-acquired MRSA, while RF122 (42) is a common strain associated with mastitis diseases in cattle.

The other three Staphylococcus species—S. epidermidis, S. haemolyticus, and S. saprophyticus—are nonvirulent but have

### Table 1. Genomes of Staphylococcus strains used in this study

<table>
<thead>
<tr>
<th>Genome (Staphylococcus spp. and putative CC for S. aureus)</th>
<th>Label</th>
<th>RefSeq accession no.</th>
<th>No. of coding sequences (total of 34,066)</th>
<th>No. of gene sets in which the genome is represented</th>
</tr>
</thead>
<tbody>
<tr>
<td>S. aureus RF122 (CC705)</td>
<td>Sau_RF122</td>
<td>NC_007622</td>
<td>2,515</td>
<td>1,178</td>
</tr>
<tr>
<td>S. aureus subsp. aureus COL (CC8)</td>
<td>Sau_COL</td>
<td>NC_002951 NC_00629*</td>
<td>2,615</td>
<td>1,205</td>
</tr>
<tr>
<td>S. aureus subsp. aureus MRSA252 (CC30)</td>
<td>Sau_MR252</td>
<td>NC_002952</td>
<td>2,656</td>
<td>1,232</td>
</tr>
<tr>
<td>S. aureus subsp. aureus MSSA476 (CC15)</td>
<td>Sau_MS476</td>
<td>NC_002953 NC_005951*</td>
<td>2,579</td>
<td>1,231</td>
</tr>
<tr>
<td>S. aureus subsp. aureus Mu50 (CC5)</td>
<td>Sau_Mu50</td>
<td>NC_00758 NC_002774*</td>
<td>2,697</td>
<td>1,242</td>
</tr>
<tr>
<td>S. aureus subsp. aureus MW2 (CC15)</td>
<td>Sau_MW2</td>
<td>NC_003923</td>
<td>2,632</td>
<td>1,238</td>
</tr>
<tr>
<td>S. aureus subsp. aureus N315 (CC5)</td>
<td>Sau_N315</td>
<td>NC_002745 NC_003140*</td>
<td>2,688</td>
<td>1,223</td>
</tr>
<tr>
<td>S. aureus subsp. aureus NCTC8325 (CC8)</td>
<td>Sau_8325</td>
<td>NC_007795</td>
<td>2,892</td>
<td>1,227</td>
</tr>
<tr>
<td>S. aureus subsp. aureus USA300 FPR3757 (CC8)</td>
<td>Sau_U300</td>
<td>NC_007793 NC_007790*</td>
<td>2,560</td>
<td>1,198</td>
</tr>
<tr>
<td>S. epidermidis ATCC 12228</td>
<td>Sep_12228</td>
<td>NC_004461 NC_005003*</td>
<td>2,419</td>
<td>1,118</td>
</tr>
<tr>
<td>S. epidermidis RP62A</td>
<td>Sep_RP62A</td>
<td>NC_002976 NC_006663*</td>
<td>2,494</td>
<td>1,115</td>
</tr>
<tr>
<td>S. haemolyticus JCSC1435</td>
<td>Shm_1435</td>
<td>NC_007168</td>
<td>2,676</td>
<td>1,141</td>
</tr>
<tr>
<td>S. saprophyticus subsp. saprophyticus ATCC 15305</td>
<td>Ssp_15305</td>
<td>NC_007350 NC_007351*</td>
<td>2,446</td>
<td>1,091</td>
</tr>
</tbody>
</table>

*The 13 genomes of Staphylococcus used in this study, along with their respective labels, RefSeq accession numbers in GenBank, the total numbers of coding sequences, and numbers of gene sets (among the total 1,354) in which the genome is represented are listed. The putative CC for each S. aureus isolate is given in parentheses in the first column.

| a | Plasmids. |

—S. epidermidis, S. haemolyticus, and S. saprophyticus—are nonvirulent but have
been associated with a number of opportunistic infections in immunocompromised patients. *S. epidermidis* strain RP62A (36) is a biofilm-forming strain that can cause toxic shock syndrome and scarlet fever, whereas strain ATCC 1228 (107) of the same species is a non-biofilm-forming, nonpathogenic strain that has been used for detecting residual antibiotics in food products. *S. haemolyticus* (strain JCSC1435) (100) and *S. saprophyticus* (strain ATCC 15305) (61) are generally opportunistic pathogens; *S. haemolyticus* infrequently causes soft tissue infections, and *S. saprophyticus* is predominantly implicated in genitourinary tract infections.

Multiple homologs within a genome can be interpreted as paralogs (a result of within-genome duplication) or xenologs (arising from acquisition of a gene copy from an external source via LGT) (33, 39, 58). Since we do not have prior knowledge on their origins, we follow Lerat et al. (64) in referring to these multiple homologs as “synologs.” We characterize the frequencies of within- and whole-gene transfer in gene sets that contain no synolog and in those that contain one or more synologs and discuss correlations with annotated gene functions. This represents the first systematic study of the transfer of genes and gene fragments in any prokaryotic genus, in which duplicated gene copies are also considered. Given that sequenced genomes can be assigned to established CCs in *S. aureus*, we also examined the extent to which LGT can disrupt established CCs.

**Materials and Methods**

**Generation of staphylococcal gene sets.** Thirteen completely sequenced genomes of *Staphylococcus* spp. were downloaded from GenBank (http://www.ncbi.nlm.nih.gov/); their accession numbers and putative CCs are presented in Table 1. These genomes represent four distinct species: *S. aureus* (nine isolates), *S. epidermidis* (two isolates), *S. haemolyticus* (one isolate), and *S. saprophyticus* (one isolate).

The 34,066 protein sequences predicted from these genomes were clustered into 2,924 sets based on similarity matching via a Markov clustering algorithm (inflation parameter 1.1) (27). These sets consist of putatively homologous protein sequences; some are encoded by one or more copies of genes within a single genome, i.e., synologs. Multiple sequence alignment was performed on each protein set using T-COFFEE (78), with four combinations of the penalties for gap opening (25 and 50) and gap extension (0 and 10), and MUSCLE (26), with the default settings. The alignments were validated by using a pattern-centric objective function (7); the alignment receiving the highest score according to the objective function was selected as optimal for each protein set.

The protein alignments were converted into nucleotide sequence alignments, with nucleotide triplets arranged to parallel exactly the protein alignment in each case. For these nucleotide sequence alignments, we require gene set sizes of 4 or greater, because 4 is the minimum size that can yield distinct topologies if every sequence in the set is unique. Where identical nucleotide sequences were present in a set, we removed (at random) all but one copy, after which gene sets with sizes of <4 (1,493; 51.1%) were again excluded. Almost all instances of the identical copies in the set is unique. Where identical nucleotide sequences were present in a set, a whole-gene transfer via LGT. To those gene sets, we applied DualBrothers (72) to define the ORBs more precisely. The program was run with an MCCM chain length of 2,500,000 generations, with a burn-in phase of 500,000 generations. The size of the window around which existing change points are randomly moved during any update was set to 5. Peter Green's constant C, the proportion of time that the sampler spends on updating the parameters when the number of dimensions is fixed, was set to 0.25. The set of trees (the tree search space) considered by the DualBrothers MCMC sampler (i.e., the list of all possible phylogenetic trees that could be inferred from shorter segments within the sequence set) was determined separately for each gene set as follows. Partitions (windows) were progressively stepped across each alignment (window length = 100, step size = 50; unit in alignment position) and the Bayesian phylogenetic inference program MrBayes (89), as described in the previous section, was used to infer unrooted phylogenetic trees at each window position. Inferred trees within the threshold at a Bayesian confidence interval of 90% were included in the initial tree list, with the maximum number of trees included set to 100. Genomes containing ORBs are inferred to have undergone one or more within-gene transfer events. (See Method S1 in reference 18 for more information regarding the generation of tree search space for the implementation of DualBrothers.)

**Detecting whole-gene (nonfragmentary) genetic transfer.** For each gene set for which no evidence of recombination was found in the first-phase screening, and for those positive in the first-phase screening but for which no recombination breakpoint was found by DualBrothers, we inferred a Bayesian phylogenetic tree and compared its topology against that of the species reference supertree using a multiple-test approach, as described in reference 15, incorporating the statistical tests of Shimodaira and Hasegawa (93), Goldman et al. (37), and Kishino and Hasegawa (57), as well as the expected likelihood weights (97), all as implemented in Tree-Puzzle 5.1 (92). Whole-gene (nonfragmentary) genetic transfer was inferred if any topology was rejected by two or more tests (P < 0.05), indicating discordance with the reference topology. The approach for analysis of genetic transfer in gene sets that contain one or more synologs was adopted from an earlier study (64). Each gene set containing m genes from a total of n distinct genomes was categorized into three separate classes: (i) those with no synolog, m = n; (ii) those with one or two synologs, n < m ≤ n + 2; and (iii) those with more than two synologs, m > n + 2. For class i gene sets with no synologs, topological discordance between a gene tree and the reference supertree implies a whole-gene transfer via LGT. For each gene set in class ii, two or four new alignment sets were generated by removing replicate synologs in all combinations, followed by tree inference and topological comparison. Discordance between one or more of these gene trees and the reference supertree implies
acquisition of a synolog in the gene set via LGT. For class iii gene sets with more than two synologs, the generation and analysis of dereplicated trees in all possible combinations quickly becomes impractical. For these sets, we compared the gene tree directly to the reference supertree; topological discordance suggests whole-gene recombination and might be explained by one or more events (see Results for details).

Indelividual gene-set trees were inferred from DNA alignments using MrBayes (89) with the parameters described above. The topology of each gene set was compared against that of the reference supertree using the multiple-test approach as described above and in reference 15. Such discordance was taken as prima facie evidence of whole-gene (nonfragmentary) genetic transfer.

Functional analysis of gene sets. Functional analysis of gene sets was based on role identifiers (Mannrole) as retrieved from the comprehensive microbial resource at The J. Craig Venter Institute (JCVI) website (http://cmr.jcvi.org/). Over- or under-representation of functional categories was based on the probability of observing a defined number of target groups (or categories) in a subsample, given a process of sampling without replacement from the whole data set (as defined in each case [see the text]) under a hypergeometric distribution (53). The probability of observing \( x \) number of a particular target category is given by the following equation:

\[
P(k = x) = \binom{m}{k} \binom{N-m}{n-k} / \binom{N}{n}
\]

where \( N \) is the total population size, \( m \) is the size of the target category within the population, \( n \) is the total size of the subsample, and \( k \) is the size of the target category within the subsample. The function for individual gene set was annotated based on sequence similarity matches (BLASTP, \( e \)-value \( \leq 10^{-5} \)) against the NCBI nr protein database, after which the annotations were extracted from the Gene Ontology database (http://www.geneontology.org/) using Blast2GO (38).

Correlation between breakpoint location and protein domain boundary. Protein domain and boundary information for each protein in the data set (\( N = 13,297 \)) was determined by sequence similarity search against domain entries (type = “domain”) in Pfam version 20 (32) and SCOP version 1.69 (1). The breakpoint-to-boundary distance \( \rho \) follows the definition in our previously published work (see Fig. 1B in reference 18) as the normalized distance from an inferred recombination breakpoint (i.e., the ORB) to the midpoint of the nearest annotated domain boundary and ranges between 0 and 1. A \( \rho \) value of \( \approx 0 \) indicates that the ORB is located at the midpoint of a domain-coding gene region (“domain”; see reference 18, i.e., no domon disruption by the ORB), whereas \( \rho = 1 \) indicates the ORB is located at or outside the domon boundary (i.e., no domon disruption by the ORB). To examine potential tendency of domon preservation during LGT (observed as large \( \rho \) values), we subsampled the data set randomly 10,000 times (omitting \( \rho = 1 \), size of each subsample = 50; to adjust for inference bias due to large sample size) and compared each subsample to a uniform distribution of (0, 1) using the Kolmogorov-Smirnov test (70).

Analysis of gene sharing relationships among Staphylococcus genomes. For each of the 34,066 proteins in our data set, we searched for the most similar-sequence (or sequences, where more than one sequence is equally most similar) in other staphylococcal isolates using BLASTP (13) against a comprehensive database comprising 212,808 protein sequences from all available 81 genomes of Staphylococcus spp. (19 complete and 62 draft genomes as of 15 March 2011 in NCBI GenBank; see Table S1 in the supplemental material). Only top hit(s) with an \( e \)-value of \( \leq 10^{-10} \) were counted. We used the term “gene affinity” to describe the relative recency of divergence (whether via vertical descent or LGT), measured as the number of matches between genomes from the same or different species, or for \( S. aureus \) the same or different CCs or species.

MLST of \( S. aureus \) isolates. We adopted the multilocus sequence typing (MLST) approach to assign CC for each of the 63 \( S. aureus \) isolates in the protein database (see Table S1 in the supplemental material). For each isolate, we identified (where possible) the housekeeping gene sequences encoding carbamate kinase (\( arcC \)), shikimate dehydrogenase (\( arcD \)), glycerol kinase (\( gfpF \)), guanylate kinase (\( gmk \)), phosphate acetyltransferase (\( pta \)), triosephosphate isomerase (\( tpi \)), and acetyl coenzyme A acetyltransferase (\( acet \)). The seven-gene allelic profile and hence the sequence type (ST) for each \( S. aureus \) isolate were identified using BLASTN (13) \( e \)-value \( \leq 10^{-10} \) against all gene alleles available from http://saureus.mlst.net/ (217 arcC alleles, 289 arcD alleles, 259 gmk alleles, 156 gmk alleles, 217 pta alleles, 225 tpi alleles, and 223 yqiL alleles as of 18 March 2011). The CC for each of the distinct STs was assigned using eBURST v3 (29) at default settings (http://saureus.mlst.net/). See File S1 in the supple-mental material for the complete allelic profiles and STs for the 63 \( S. aureus \) isolates, the distinct STs used for eBURST analysis, and the resulting output.

**RESULTS**

Of the initial 2,924 clustered protein sets, 77 (2.6%) have a size of >52 and on that basis were excluded from analysis of LGT. Many of these large sets consist of proteins related to transport functions; many of the largest is composed of 1,777 proteins (constituting 5.2% of all annotated protein-coding genes annotated in Staphylococcus) related to the ATP-binding cassette (ABC) transporter. ABC transporters constitute one of the largest sets more generally in prokaryotes (43). The second-largest set is a 769-member cluster of proteins related to the phosphotransferase system (PTS) that are involved in sugar phosphorylation and regulation of metabolic processes (90). To analyze modularity of genetic transfer (at gene instead of protein level) in these genomes, we recovered protein-coding gene sets that correspond to the protein sets. By imposing a size restriction (4 \( \leq N \leq 52 \)) and removing identical sequences (see Materials and Methods), we retained a total of 1,354 protein-coding gene sets in the subsequent analysis (4 \( \leq N \leq 41 \); no gene set fell in the range 42 \( \leq N \leq 52 \)). The representation of each genome among the 1,354 gene sets and the size distribution for these sets are shown in Table 1 (far-right column) and Fig. 1, respectively. The most virulent species, \( S. aureus \), is well represented, with all but two \( S. aureus \) genomes represented in over 1,200 gene sets. In total, for 229 (15.7%) gene sets, \( N = 8 \), and for 197 (14.5%), \( N = 9 \). The largest set consists of 41 sequences that encode putative nucleotidase proteins, while the second-largest set has 39 putative proteins of guanosine- or inositol-monophosphate dehydrogenase.

**Inference of genetic transfer.** We inferred genetic transfer differently in single-versus multicopy gene sets. Single-copy (SC) gene sets contain no within-genome duplicated genes, i.e., no genome is represented more than once in such a set. In the absence of evidence to the contrary, the most parsimonious interpretation is that these genes share an evolutionary origin that predates the divergence of these strains, i.e., that these genes are orthologs. On the other hand, multicopy (MC) gene sets contain one or more within-genome duplicates, with at
least one genome represented more than once in such a set. The additional gene copy (it is neither necessary, nor indeed usually possible, to distinguish the “original” from the “duplicate”) must either have arisen via a within-genome duplication event (i.e., be a paralog) or have been imported via LGT into the genome (i.e., be a xenolog). The presence or absence of these synologs necessarily affects how we interpret evidence of genetic transfer, as discussed below.

We interpret the evidence of phylogenetic discrepancies (manifested as topological discordance) and the inference of one or more observed recombination breakpoints, i.e., ORB(s), within the boundaries of the gene set as within-gene (fragmentary) genetic transfer (i.e., instances of ORB⁺; see reference 15). Table 2 shows the possible interpretations of fragmentary transfer in SC and MC gene sets. For those sets in which we find no ORB (instances of ORB⁻), we infer that there has been no fragmentary genetic transfer (SC-noFragGT and MC-noFragGT); in the latter case (MC), the provenance of the synologs is unclear. Inference of ORBs within SC gene sets (SC-FragGT) can be interpreted (again, in the absence of evidence to the contrary) as LGT between orthologs. In MC gene sets, the situation is again more complex: (i) if LGT is inferred within the set but the recombinants are xenologs (not chimeras, since these genes have been transferred in their entirety), and we denote this situation in the set by MC-WholeGT-PX with the same qualification as in the case of within-gene transfer. If (iii) all synologs are recombinant, we infer that each is a xenolog (again, more precisely, a paralog-xenolog chimera), but it may be impossible to reconstruct the precise order of within-genome duplication, LGT and perhaps other (e.g., gene conversion or lineage-sorting) events that have produced this situation, which we denote as MC-FragGT-PX.

The interpretation is much the same in the case of whole-gene transfer, as shown in Table 3 for SC and MC gene sets, although with a complication (discussed below) that arises from our methodological approach. In the case of SC gene sets, both alternatives (SC-noWholeGT and SC-WholeGT) exactly parallel those for within-gene transfer (Table 2), with whole-gene transfer in SC gene sets again interpreted as involving orthologs. In MC gene sets, the situation is again more complex: (i) if LGT is inferred within the set but the synologs are not implicated, then all synologs are paralogs (MC-WholeGT-P), but (ii) where some but not all of the synologs are recombinant, the nonrecombinant synologs are native paralogs but the recombinants are xenologs (not chimeras, since these genes have been transferred in their entirety), and we denote this situation in the set by MC-WholeGT-P/X. Where per-genome copy number is small, it is reasonable to assume (in the absence of evidence to the contrary) that the LGT event occurred subsequently to the within-genome duplication. If, however, (iii) all synologs are detected as recombinant, we know that each is a xenolog (again, more precisely, a paralog-xenolog chimera), but it may be impossible to reconstruct the precise order of within-genome duplication, LGT and perhaps other events that have produced this situation, which we denote as MC-WholeGT-PX.

The complication, mentioned above, concerns the evidentiary basis on which we identify none, some, or all synologs as recombinant via whole-gene transfer. In our approach, we inferred an ORB as a point at which phylogenetic trees inferred the precise order of within-genome duplication, LGT and perhaps other events that have produced this situation, which we denote as MC-WholeGT-PX.

### Table 2. Inference based on evidence of within-gene transfer in single-copy and multicopy gene sets

<table>
<thead>
<tr>
<th>Gene set</th>
<th>Negative</th>
<th>Positive (ORB⁺ gene set)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SC gene set</td>
<td>No LGT; SC-noFragGT</td>
<td>Within-gene LGT between orthologs SC-FragGT</td>
</tr>
<tr>
<td>MC gene set</td>
<td>No LGT; MC-noFragGT</td>
<td>(i) No synologs are recombinant (synolog = paralog) MC-FragGT-P</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(ii) Some synologs are recombinant (synolog = paralog or xenolog) MC-FragGT-P/X</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(iii) All synologs are recombinant (synolog = paralog and/or xenolog) MC-FragGT-PX</td>
</tr>
</tbody>
</table>

* The inference was based on evidence of within-gene transfer in single-copy (SC) and multicopy (MC) gene sets, for cases with negative evidence (noFragGT) and positive evidence inferred as LGT (FragGT). The gene sets denoted by FragGT are the ORB⁺ sets. In the MC gene sets, the suffix -P indicates that synologs are paralogs, and -X indicates that synologs are xenologs. The suffix -PX denotes that each synolog is either a paralog or a xenolog (but not both), while -PX indicates that each synolog can have a complex history, i.e., be a paralog, a xenolog, or both. See the text for details.

### Table 3. Inference based on evidence of whole-gene transfer in single-copy and multicopy gene sets

<table>
<thead>
<tr>
<th>Gene set</th>
<th>Evidence of whole-gene transfer</th>
</tr>
</thead>
<tbody>
<tr>
<td>SC gene set</td>
<td>No LGT; SC-noWholeGT</td>
</tr>
<tr>
<td>MC gene set</td>
<td>No LGT; MC-noWholeGT-P</td>
</tr>
<tr>
<td></td>
<td>Whole-gene LGT involving orthologs SC-WholeGT</td>
</tr>
<tr>
<td></td>
<td>(i) No recombining sequences are synologs (synolog = paralog) MC-WholeGT-P</td>
</tr>
<tr>
<td></td>
<td>(ii) Some recombining sequences are synologs (synolog = paralog or xenolog) MC-WholeGT-P/X</td>
</tr>
<tr>
<td></td>
<td>(iii) All recombining sequences are synologs (synolog = paralog and/or xenolog) MC-WholeGT-PX</td>
</tr>
</tbody>
</table>

* Inference was based on evidence of whole-gene transfer in single-copy (SC) and multicopy (MC) gene sets, for cases with negative evidence (noWholeGT) and positive evidence inferred as LGT (WholeGT). Gene sets denoted by WholeGT are the ORB⁺ sets. The label WholeGT distinguishes these cases from the within-gene transfer (FragGT) shown in Table 2. Labels otherwise follow the conventions introduced in Table 2. See the text for details.
alignments are concordant, suggesting that both phylogeny. There are four possible outcomes in such a comparison with respect to the reference phylogeny: (a) both trees of the de-replicated only once. A Bayesian phylogenetic tree was constructed for each of these alignments, and the tree was compared against the reference species gene copies, \( a_1 \) suggesting that \( a_1 \) a1

of alignments in which each genome is represented only once. A Bayesian phylogenetic tree was constructed for each of these alignments, and the tree was compared against the reference species phylogeny. There are four possible outcomes in such a comparison with respect to the reference phylogeny: (a) both trees of the de-replicated alignments are concordant, suggesting that both \( a_1 \) and \( a_2 \) are paralogs; (b and c) one of the trees is concordant, while the other is discordant, suggesting that \( a_1 \) and \( a_2 \) have different evolutionary trajectories, i.e., \( a_1 \) is a paralog, \( a_2 \) is a xenolog, or vice versa; and (d) both trees are discordant, suggesting that both \( a_1 \) and \( a_2 \) have a complex history, i.e., these copies are xenologs following gene duplication or paralogs following gene transfer, of which the order of the events is undetermined. The labels for each inference of synology are shown for multicopy gene sets (MC) and in cases of within-gene transfer (FragGT) and whole-gene transfer (WholeGT). (Modified from Fig. 2 in reference 64.)

FIG. 2. Testing the source of synologs in multicopy gene sets. The illustrated example is a set with a single synolog, in which genome \( A \) has two gene copies, \( a_1 \) and \( a_2 \). The gene copies are deroeplicated in all combinations, yielding two sets of alignments in which each genome is represented only once. A Bayesian phylogenetic tree was constructed for each of these alignments, and the tree was compared against the reference species phylogeny. There are four possible outcomes in such a comparison with respect to the reference phylogeny: (a) both trees of the de-replicated alignments are concordant, suggesting that both \( a_1 \) and \( a_2 \) are paralogs; (b and c) one of the trees is concordant, while the other is discordant, suggesting that \( a_1 \) and \( a_2 \) have different evolutionary trajectories, i.e., \( a_1 \) is a paralog, \( a_2 \) is a xenolog, or vice versa; and (d) both trees are discordant, suggesting that both \( a_1 \) and \( a_2 \) have a complex history, i.e., these copies are xenologs following gene duplication or paralogs following gene transfer, of which the order of the events is undetermined. The labels for each inference of synology are shown for multicopy gene sets (MC) and in cases of within-gene transfer (FragGT) and whole-gene transfer (WholeGT). (Modified from Fig. 2 in reference 64.)
Gene sets affected by fragmentary transfer are significantly over-represented, compared to expectation, in various categories (energy metabolism; DNA metabolism; protein fate; amino acid biosynthesis; purines, pyrimidines, nucleosides, and nucleotides; and transcription) for both SC-FragGT and MC-FragGT. On the other hand, both types of gene sets are significantly under-represented in the categories hypothetical protein, transport and binding proteins, and regulatory functions. The category protein synthesis is over-represented in SC-FragGT but under-represented in MC-FragGT, while the unknown function category is biased in the opposite direction.

**Whole-gene (nonfragmentary) genetic transfer.** Of the 1,354 staphylococcal gene sets under consideration, we have thus far inferred within-gene transfer for 320 (252 as clear instances of LGT and a further 68 for which evidence was deemed inconclusive). We now turn to the 1,034 remaining gene sets: 953 recombination-negative plus 81 false-positive cases resulting from our two-phase approach for detecting within-gene transfer. For each we inferred a Bayesian phylogenetic tree and compared it with the reference MRP supertree generated from 2,645 putatively orthologous sets in these 13 genomes (see Materials and Methods). The MRP supertree is shown in Fig. 4a. Four *Staphylococcus* species are represented among the 13 genomes, each monophyletic (*S. saprophyticus* and *S. haemolyticus* trivially so) according to our supertree analysis. Using the MLST technique, we grouped the nine *S. aureus* isolates into five distinctive CCs: (i) the strains MW2 and MSSA476 in CC15, (ii) COL, USA300 and NCTC8325 in CC8, (iii) N315 and Mu50 in CC5, (iv) MRSA252 in CC30, and (v) RF122 in CC705. In our analysis, STs previously identified as CC1, i.e., those of MW2 and MSSA476, were grouped as CC15, likely due to the increased frequency of ST15 in the current database (http://saureus.mlst.net/); the naming convention of CC is...
subunit rRNA genes (99) that gives *S. saprophyticus* phage transcriptional activator (gene set 267), an instance of (the synolog is either a paralog or a xenolog, but not both) (c); and 

\[ \text{MC-WholeGT-P} \]

mycin resistance protein/dioxygenases (gene set 445), an instance of (a), and tree topologies for gene sets encoding for glyoxalase/bleomycin resistance protein/dioxygenases isomerase (gene set 1167), an instance of \[ \text{MC-WholeGT-P/X} \].

Staphylococcus species using small subunit rRNA genes (99). Whole-gene transfer was inferred in a gene set if the tree topology was significantly discordant with that of the reference supertree (see Materials and Methods).

The 1,034 gene sets can be divided into three classes based on the number of synologs present in each set. The number of gene sets with trees concordant or discordant with the reference supertree for each category is shown in Table 4. Among the 774 of these sets in SC, 97 (13%) show topological discordance and represent LGT involving orthologs (SC-WholeGT), whereas no evidence of LGT was found among the other 677 (87%), i.e., SC-noWholeGT. Among the 260 of these sets in MC, however, the outcome was very different: 210 (81%) are discordant (MC-WholeGT) vis-à-vis the reference topology and only 50 (19%) concordant (MC-noWholeGT). Of these 260 sets, 105 contain one or two synologs each, and 155 contain more than two. Of the 105 sets in the former group, 58 (55%) revealed evidence of whole-gene transfer, whereas, remarkably, among the 155 in the latter group, fully 152 (98%) did so.

Using our synolog dereplication approach (Fig. 2), we could further classify the 58 MC-WholeGT sets with one or two synologs into 12 for which all synologs are paralogs but not xenologs (MC-WholeGT-P), 22 for which synologs are either paralogs or xenologs but not both (MC-WholeGT-PX), and 19 with more-complex histories (MC-WholeGT-PX). The remaining five were reduced to gene sets with sizes of <4 upon our dereplication approach and thus do not provide meaningful topology comparison under this classification. Examples of tree topologies for each of these instances are shown in Fig. 4b through Fig. 4d; for each tree, a homologous gene copy from *Bacillus cereus* or *B. thuringiensis* was used as outgroup. Figure 4b shows the topology for staphylococcal genes encoding the function annotated as glyoxalase/bleomycin resistance protein/dioxygenases. In comparison to the reference supertree in Fig. 4a, the intraspecies gene sharing among lineages of *S. aureus* (not any of the gene copies in *S. saprophyticus*) contributed to the topological incongruence between the two trees, i.e., both the *S. saprophyticus* synologs are paralogs (an instance of MC-WholeGT-P). On the other hand, gene copies of *S. saprophyticus* encoding glucosamine-6-phosphate isomerase (Fig. 4c) show a different pattern (an instance of MC-WholeGT-PX), in which a copy shows evidence of gene sharing with lineages of *S. epidermidis* or *S. aureus* (i.e., a xenolog), whereas another based on the predicted ST founder, i.e., the most frequent ST (ST15 instead of ST1) found within the group. The supertree is rooted using *S. saprophyticus* as outgroup, based on a previous phylogenetic study of *Staphylococcus* species using small subunit rRNA genes (99). Whole-gene transfer was inferred in a gene set if the tree topology was significantly discordant with that of the reference supertree (see Materials and Methods).

The frequency of concordance and discordance in these sets, based on four maximum-likelihood tests, are shown; the proportions are relative to the total data set of 1,354 gene sets. Labels of different categories follow the description in Table 3. SC, single copy; MC, multicopy.

### Table 4. Counts of inferred whole-gene transfer based on topological comparison between the phylogenetic tree inferred for each gene set and the reference supertree

<table>
<thead>
<tr>
<th>Category</th>
<th>Tree comparison based on maximum-likelihood tests</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Concordant</td>
</tr>
<tr>
<td>SC gene sets, no synologs</td>
<td>SC-noWholeGT, 677 sets</td>
</tr>
<tr>
<td>Gene sets with one or two synologs each (within MC gene sets)</td>
<td>MC-noWholeGT-P, 47 sets</td>
</tr>
<tr>
<td>Gene sets with more than two synologs each (within MC gene sets)</td>
<td>MC-noWholeGT-P, 3 sets</td>
</tr>
<tr>
<td>Total no. of gene sets (%)</td>
<td>727 (53.7)</td>
</tr>
</tbody>
</table>

*The frequency of concordance and discordance in these sets, based on four maximum-likelihood tests, are shown; the proportions are relative to the total data set of 1,354 gene sets. Labels of different categories follow the description in Table 3. SC, single copy; MC, multicopy."
has likely arisen by gene duplication (i.e., is a paralog). For the
gene set encoding phage transcriptional activator (Fig. 4d), an
important protein implicated in phage-mediated transduction,
the evolutionary history of both \textit{S. aureus} NCTC8325 gene
copies could have involved duplication and/or LGT, of which
the order of the events could not be determined using this
approach (MC-WholeGT-PX).

Figure 5 shows the proportions of proteins in each JCVI
functional category within (Fig. 5a) single-copy (SC-WholeGT)
and (Fig. 5b) multicopy (MC-WholeGT) gene sets for which we
inferred whole-gene transfer, compared to their frequencies in
the full staphylococcal data set. The proteins in these sets are
significantly either over- or under-represented in more than
half of the JCVI categories, although with more-numerous and
greater differences between SC and MC gene sets than was
observed for fragmentary transfer. SC sets affected by whole-
gene transfer are significantly over-represented, compared to
expectation, in the protein synthesis, central intermediary me-
tabolism, and transcription categories (in contrast to under-
representation of these same categories in MC sets), suggest-
ing that \textit{Staphylococcus} genome lineages appear to have been
more receptive to introgression of whole genes that encode for
these protein functions when no indigenous copy was already
present, or if an indigenous copy was present it was replaced or
subsequently lost.

Correspondingly, gene sets annotated with these same func-
tional categories are significantly under-represented in
MC-WholeGT, suggesting that these genomes have been less recep-
tive to the integration of genes encoding these functions when
multiple copies already exist. The category cellular processes is
over-represented in SC-WholeGT, particularly the functions
involved in toxin production and resistance (\(P = 1.7 \times 10^{-4}\)),
while the categories implicated in metabolic functions, includ-
ing energy metabolism, are under-represented. The categories
cell envelope and mobile and extrachromosomal genetic ele-
ment functions are over-represented in MC-WholeGT. The

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig5}
\caption{Representation of functional categories assigned to protein sequences corresponding to gene sets in \textit{Staphylococcus} species that show
evidence of whole-gene transfer (\(\square\)) for single-copy gene sets (SC-WholeGT) (a) and multicopy gene sets (MC-WholeGT) (b). The solid bars (\(\textcolor{red}{\square}\))
show the same functional categories in the full data set (1,354 sets, 13,297 proteins). Categories are numbered differently for panels a and b shown
in the boxes. Functional categories that are over-represented in panel a but under-represented in panel b, or vice versa, are indicated in boldface.
The significance of over- or under-representation is represented by single (\(P \leq 0.05\)) and double (\(P \leq 0.01\)) asterisks.}
\end{figure}
categories fatty acid and phospholipid metabolism and signal transduction are under-represented in both SC-WholeGT and MC-WholeGT.

A summary of the results for all 1,354 gene sets used here, including their putative functional annotations, is available in File S2 in the supplemental material. The (protein and nucleotide) alignments and the Bayesian phylogenies for all 1,354 gene sets generated from the present study are available at http://bioinformatics.org.au/staphGT/. Of 559 gene sets that show evidence of LGT (both FragGT and WholeGT), 146 (26.1%) were not annotated with a function, while 277 (49.5%) have functions related to enzymatic reactions, e.g., kinases, isomerases, and transferases. The majority of genes related to resistance to antibiotics, drugs or heavy metals (11 of 15) are found to have undergone LGT; seven of these are in WholeGT, including genes conferring resistance to penicillin, methicillin, and heavy metals, including aluminum and arsenic (see Table S2 in the supplemental material). Of the three gene sets with annotated functions related to bacterial toxins, we found evidence of whole-gene transfer in the gene set encoding Txe family addiction module toxin but not in the other two, exfoliative and MazF toxins. In another group are gene sets showing evidence of LGT and annotated with functions related to membrane transport (39 of 59; see Table S3 in the supplemental material). Of the three gene sets with annotated functions related to bacterial toxins, we found evidence of whole-gene transfer in the gene set encoding Txe family addiction module toxin but not in the other two, exfoliative and MazF toxins. In another group are gene sets showing evidence of LGT and annotated with functions related to membrane transport (39 of 59; see Table S3 in the supplemental material).

Interestingly, among the 60 sets of ribosomal proteins in the 1,354 data set (and their related subunits, which are expected to be highly conserved), we found evidence of FragGT in six gene sets and WholeGT in another four (two of which are related to methyltransferases of rRNA small subunits). This is in addition to LGT evidence in 14 of 27 gene sets (9 of which are in WholeGT) that are related to transcriptional regulatory functions, e.g., transcription activators and elongation factors.

Correlation of transferred genomic regions with protein structural domains. Within each ORB gene set, we further sought to determine whether the transferred coding regions correlate with protein structural domains. Using domain information available from the Pfam (32) and SCOP (1) databases, we introduced the ρ statistic to examine the potential tendency of LGT to disrupt domain-coding gene region, i.e., domon (see reference 18 and Materials and Methods). A ρ value of ~1 indicates that the ORB is located at or outside the domon boundary. If there is a strong correlation between domon boundaries and ORB, the observed ρ values would be significantly larger than values uniformly distributed at random. We compared the distribution of observed ρ values against a uniform distribution (between 0 and 1), taking potential bias in large sample sizes into account via a random subsampling approach (10,000 subsamples of 50 values each, see Materials and Methods). Figure 6 shows the results of this analysis for SC gene sets (Fig. 6a), i.e., SC-FragGT and MC gene sets (Fig. 6b), i.e., MC-FragGT. For each, we show the observed distribution of ρ values (subpanel i) and, across the 10,000 Kolmogorov-Smirnov tests between an observed and an expected distribution, the distribution of D (subpanel ii) and P (subpanel iii) values. The D values indicate the magnitude of difference between the observed and expected distributions, whereas the P values indicate the statistical significance of such difference. Both of these values range between 0 and 1. In this instance, small P values indicate that the observed ρ values are larger than expected under a uniform distribution.

For the cases of SC-FragGT and MC-FragGT, any bias in the observed distribution of ρ values is unclear, as shown in panels i in Fig. 6. For SC-FragGT (Fig. 6a) we found small differences between observed and expected ρ values (Kolmogorov-Smirnov test, mean D value = 0.16; ca. 75% of the P values > 0.1). A similar trend was observed in MC-FragGT (Fig. 6b), with most (ca. 92%) of the observed P values > 0.1 (mean D value = 0.11).

Staphylococcal MGEs and LGT. We identified 37 gene sets among the 1,354 that encode functions known to be present in staphylococcal mobile genetic elements (MGEs) (see Table S4 in the supplemental material) as reported in previous studies (66, 69). Among these 37 gene sets, 31 (83.8%) show clear evidence of LGT (18 instances of WholeGT and 11 instances of FragGT); for 2 the evidence was inconclusive, and for 6 we found no evidence of LGT. Although functions related to MGE are significantly (P ≤ 0.01) over-represented in MC-WholeGT (Fig. 5b) and most MGE genes show high suscepti-
bility to LGT, most of the gene sets we identified as affected by LGT are annotated with functions that extend beyond virulence or resistance to include a broad scope of central metabolism, transcription, and translation (289 instances in WholeGT and 241 instances in FragGT; see File S2 in the supplemental material).

Genome affinities of \textit{S. aureus} clonal complexes and staphylococcal species. Figure 7 shows the genome affinity for each of the 13 staphylococcal genomes, assessed by the best BLAST matches (most-similar protein sequences) for each of these 34,066 proteins in other staphylococcal isolates/species (see also Table S5 in the supplemental material). Most proteins (>85% across all genomes) from \textit{S. aureus} and \textit{S. epidermidis} find a best match or matches within the same species, suggesting high level of intraspecies genome affinity. \textit{S. haemolyticus} and \textit{S. saprophyticus} could not be evaluated in this way, since they are lone representatives of their respective named species among these complete and draft genomes. For \textit{S. aureus}, we further examined whether these proteins are most similar to sequences from other isolates in the same CC or in another CC. Across the nine \textit{S. aureus} genomes (constituting five different CCs), most proteins find their best matches outside their own CC: 57.9% (of total proteins in a genome) on average between the two CC5 isolates, 62.7% on average among the three CC8 isolates, 67.6% on average between the two CC15 isolates, and 77.3 and 96.5% for each of the genomes representing CC30 and CC705, respectively (see Table S5 in the supplemental material). This suggests a high level of inter-CC genome affinity among \textit{S. aureus} isolates. CC5, CC8, and CC15 mutually share high genome affinities, since each is among the three CCs most similar to each of the other two (see Table S5 in the supplemental material). Genomes of CC30 are highly similar to those of CC30, CC42, and CC8, and those of CC705 are highly similar to those of CC8, CC10, and CC5, although for CC30 and CC705 in particular these affinities may represent rough approximations due to the limited data currently available.

DISCUSSION

Two-thirds of the strains of \textit{S. aureus} in the present study (all except MSSA476, NCTC8325, and RF122) are methicillin resistant. According to the MRP supertree based on sets of protein-coding genes from the 13 genomes examined here (Fig. 4a), the methicillin-susceptible isolates nest within different clades in the supertree admixed with resistant isolates, implying multiple origins of methicillin resistance in \textit{S. aureus} (20).

Among 1,354 sets of homologous genes, we found clear evidence of LGT in 368 gene sets (252 ORB sets in SC-FragGT and MC-FragGT, 97 sets in SC-WholeGT, and 19 sets in MC-WholeGT-PX, totaling 27.1% of 1,354). The remaining 191 sets in MC-WholeGT, together with the 68 inconclusive cases from analysis of within-gene transfer (totaling 259; 19.1% of 1,354), represent probable LGT, although the exact origins of synologs in these instances could not be determined. Leaving aside the 68 inconclusive cases, within-gene (252; 18.6%) and whole-gene (307; 22.7%) transfer contribute almost equally to the discordance of these gene sets against the reference species phylogeny. We observe a higher frequency of inferred genetic transfer involving multiple (MC) than single-copy (SC) gene sets, some of which can be explained by LGT.
alone, but most of which appears to reflect more-complex evolutionary histories involving multiple LGT and gene duplication events, gene conversion, and/or lineage sorting. On the other hand, no phylogenetic approach can detect recombination between genomic lineages that are terminal and adjacent on a given gene-set tree, suggesting that our estimates of recombination frequency are probably low.

LGT and gene duplication have been proposed to be the major factors contributing to functional innovation in prokaryotes (68, 103). Hooper and Berg (46) proposed that duplication may be more common among laterally transferred genes than among indigenous ones. Although our study was not explicitly designed to test this hypothesis, based on the results presented here we cannot reject this assertion, since we found evidence of LGT among a larger proportion of MC genes (364/433; 84.1%) than SC gene sets (195/921; 21.2%); see File S2 in the supplemental material. In addition, of the 11 gene sets related to antibiotic/drugs or heavy metal resistance that show evidence of LGT (see Table S2 in the supplemental material), 9 (81.8%) are MC. In contrast, gene sets conferring similar functions that show no evidence of LGT are mostly SC (four of five). Our observations suggest that both gene duplication and LGT are important in the maintenance of the resistance mechanisms to toxic drugs or heavy metals in Staphylococcus. This finding is in agreement with a recent study that demonstrates gene amplification as a key factor in reducing fitness costs associated with antibiotic resistance in bacteria (76). Assuming that maintenance of resistance mechanisms is deleterious in the absence of target molecules (e.g., antibiotics, drugs, or heavy metals), the dosage compensation of these (largely externally acquired) narrow-function genes by duplication is likely a regulatory response to maintain resistance (91).

Gene sets for which we infer LGT are represented at frequencies significantly different from random expectation in more than half of the JCVI functional role categories. The difference in the patterns of over- and under-representation between SC and MC gene sets is greater for whole-gene than for fragmentary genetic transfer. Gene sets involved in protein synthesis and affected by LGT, whether within-gene or whole-gene, are highly over-represented among SC sets but under-represented among MC sets, suggesting that these Staphylococcus genomes are more susceptible to introgression via LGT of genes related to amino acid synthesis when no similar copy is already present in the recipient genome than when multiple copies already exist. For whole-gene transfer, we likewise found significant bias in favor of SC sets in the frequency of protein sets engaged in toxin production and resistance. Genes related to toxin-antitoxin systems have recently been implicated in LGT between Staphylococcus and Listeria (19). The system is thought to be a selective mechanism for postsegregational killing of daughter cells that lack the genes, usually by cleaving or modifying nucleotides or ribosomes (101), e.g., the control of endoribonuclease activity by the well-studied MazFE system (25). Both genes encoding MazF toxin and MazE antitoxin are recovered as SC gene sets in our 1,354-set data. These two proteins are usually present together, forming a linear heterodimer structure (35, 54). We found evidence of whole-gene transfer implicating the MazF toxin in Staphylococcus, but not in the corresponding antitoxin MazE, suggesting different evolutionary histories between the two and that the antitoxin MazE is largely inherited vertically among the genus. Alternatively, highly similar (>90% nucleotide identity) sequences of MazE in Staphylococcus (e.g., due to convergence) could have obscured the LGT signal from detection using our approach.

MGEs have been reported to be the key agents of LGT that contribute to the rapid spread of virulence and antibiotic resistance in Staphylococcus (9, 65, 66, 69). Our findings, based on a subset of 1,354 gene sets, demonstrate clear evidence of LGT among staphylococci that extends beyond functions related to MGE and virulence. We found LTG to be more frequent among different CCs within S. aureus than between different staphylococcal species. Interestingly, of seven housekeeping genes represented in our 1,354 gene sets and commonly used for MLST analysis, three (glycerol kinase [gkpF], phosphate acetyltransferase [pta], and triosephosphate isomerase [tpi]) show clear evidence of fragmentary transfer, implicating LGT even in the most conserved genes in the genus and again pointing to a footprint of LGT that extends much farther beyond MGE functions than previously appreciated (19, 96). This level of mutual gene sharing identifies genus Staphylococcus as a genetic exchange community (95), within which barriers against LGT appear to be low.

The complexity hypothesis (51) postulates that genes involved in functions related to complex regulatory networks such as transcription and translation (i.e., genes encoding informational proteins) are less likely to undergo LGT compared to genes encoding operational (e.g., metabolic) functions. Our results do not speak directly to this hypothesis but, interestingly, show that some of the most conserved genes that encode informational proteins (ribosomal proteins and transcription factors) are susceptible to both within-gene and whole-gene LGT in Staphylococcus. The introgression of these genetic fragments into the bacterial genome could be an important repair process to preserve gene function.

Almost all previous approaches to quantifying LGT in prokaryotes, including all those based on a phylogenetic approach (8, 63, 64, 85, 108), have been based on the assumption that the unit of genetic transfer is necessarily an entire (full-length) gene. Here, for within-gene genetic transfer in Staphylococcus genomes, we found only a very weak correlation between inferred breakpoints and domon boundaries. This observation is in agreement with a recent study (18) in which domons were found to not have been preferentially preserved intact during LGT among prokaryotes more broadly. This Staphylococcus data set, or indeed any within which homologous genes have diverged relatively little, may not be ideal for the phylogenetic exploration of possible relationships between recombined regions and domons, since the signal of recombination (or LGT) would not have been immediately obvious. Highly similar sequences are unlikely to be disruptive of protein domain structure, even when the recombination breakpoint is internal to a domon, and breakpoints become increasingly difficult to locate precisely at high similarity between introgressed and native sequences, as characteristic differences necessarily become less frequent.

Both LGT (whether of partial or complete genes) and genetic duplication shape the functional evolution of protein families in eukaryotes (3). Our results demonstrate that both
processes have contributed to gene diversity and functional innovation among staphylococcal genomes. Indeed, since our approach cannot detect LGT between immediate sister genomes at the tips of the species tree (e.g., between lineages of highly similar \textit{S. aureus} CCs), our results almost certainly underestimate the extent of genetic transfer in these genomes. Other limitations of the phylogenetic approach in delineating LGT have been described in detail elsewhere (5, 16, 59, 88).

We have also identified a substantial class of genes for which there is evidence of a complicated evolutionary history that includes multiple events of genetic transfer and/or genetic duplication, e.g., recombination between two synologs or duplication of synologs. These results significantly enhance our understanding of genome evolution, particularly genetic transfer, in the genus \textit{Staphylococcus}, including virulently pathogenic isolates.

**ACKNOWLEDGMENTS**

This study was supported by Australian Research Council grant CE1038221.

We thank Aaron Darling and Vladimir Minin for valuable advice on the use of DualBrothers.

**REFERENCES**


