JOURNAL OF BACTERIOLOGY, Aug. 2011, p. 3964–3977 Vol. 193, No. 15
0021-9193/11/$12.00 doi:10.1128/JB.01524-10
Copyright © 2011, American Society for Microbiology. All Rights Reserved.

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

Cheong Xin Chan,[1]‡ Robert G. Beiko,[2] and Mark A. Ragan[1]*

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, Canada[2]

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, septicemia (bacteremia), 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 SCC*mec* 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-

* Corresponding author. Mailing address: ARC Centre of Excellence in Bioinformatics and Institute for Molecular Bioscience, The University of Queensland, Brisbane, Queensland 4072, Australia. Phone: 61-7-33462616. Fax: 61-7-33462101. E-mail: m.ragan@uq.edu.au.

† Supplemental material for this article may be found at http://jb.asm.org/.

‡ Present address: Department of Ecology, Evolution, and Natural Resources, Rutgers University, New Brunswick, NJ 08901.

▽ Published ahead of print on 27 May 2011.

Downloaded from http://jb.asm.org/ on March 30, 2015 by guest

TABLE 1. Genomes of *Staphylococcus* strains used in this study[a]

Genome (*Staphylococcus* spp. and putative CC for *S. aureus*)	Label	RefSeq accession no.[b]	No. of coding sequences (total of 34,066)	No. of gene sets in which the genome is represented
S. aureus RF122 (CC705)	Sau_RF122	NC_007622	2,515	1,178
S. aureus subsp. *aureus* COL (CC8)	Sau_COL	NC_002951	2,615	1,205
		NC_006629*	3	
S. aureus subsp. *aureus* MRSA252 (CC30)	Sau_MR252	NC_002952	2,656	1,232
S. aureus subsp. *aureus* MSSA476 (CC15)	Sau_MS476	NC_002953	2,579	1,231
		NC_005951*	19	
S. aureus subsp. *aureus* Mu50 (CC5)	Sau_Mu50	NC_002758	2,697	1,242
		NC_002774*	34	
S. aureus subsp. *aureus* MW2 (CC15)	Sau_MW2	NC_003923	2,632	1,238
S. aureus subsp. *aureus* N315 (CC5)	Sau_N315	NC_002745	2,688	1,223
		NC_003140*	31	
S. aureus subsp. *aureus* NCTC8325 (CC8)	Sau_8325	NC_007795	2,892	1,227
S. aureus subsp. *aureus* USA300 FPR3757 (CC8)	Sau_U300	NC_007793	2,560	1,198
		NC_007790*	5	
		NC_007791*	3	
		NC_007792*	36	
S. epidermidis ATCC 12228	Sep_12228	NC_004461	2,419	1,118
		NC_005003*	11	
		NC_005004*	22	
		NC_005005*	16	
		NC_005006*	8	
		NC_005007*	6	
		NC_005008*	3	
S. epidermidis RP62A	Sep_RP62A	NC_002976	2,494	1,115
		NC_006663*	32	
S. haemolyticus JCSC1435	Shm_1435	NC_007168	2,676	1,141
S. saprophyticus subsp. *saprophyticus* ATCC 15305	Ssp_15305	NC_007350	2,446	1,091
		NC_007351*	45	
		NC_007352*	23	

[a] 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.

[b] *, Plasmids.

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

Downloaded from http://jb.asm.org/ on March 30, 2015 by guest

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, 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 removed from the data set (99.9% of 7,356 sequences) represented organisms annotated as the same species or strain; therefore, the effect of this step in altering final tree topologies among distinct genome isolates (in subsequent analysis [see below]) is negligible. In addition, we excluded gene sets with sizes of >52 (77; 2.6%) for computational reasons: Bayesian phylogenetic inference is computationally very demanding for large sequence sets (56 of which, i.e., 1.9%, have sizes of >70), and interpretation of the resulting tree is often problematic owing to poor convergence and the multiplicity of possible topology resolution paths. In this way, the data set used subsequently in the present study was reduced to 1,354 sets ($4 \leq N \leq 52$) corresponding to 13,297 genes (39.0% of all genes annotated in these genomes; the representation of the genomes in these gene sets is shown in Table 1). Each of the 13 genomes, if equally represented in a gene set of size 52, can have a maximum of four gene copies. The overall pairwise nucleotide sequence similarity across all gene sets ranges from 0.47 to 0.99, with a mean of 0.74 and standard deviation of 0.12.

Reference species tree. As a reference phylogenetic tree for the 13 *Staphylococcus* isolates, we computed a supertree using the matrix representation with parsimony (MRP) approach (87). Hybrid clustering (40) of all proteins based on their pairwise distances (BLASTP; *e*-value $\leq 10^{-3}$) among 34,066 protein-coding sequences yielded 2,645 putatively orthologous protein sets (maximally representative clusters) (8). Sequences in each set were aligned as described above and phylogenetic trees were inferred using MrBayes (89) with an MCMC chain length of 2,500,000, of which the first 500,000 generations, comprising 5,001 sampled trees, were discarded; the K2P nucleotide substitution model (56); and a four-category approximation to the gamma distribution for among-site rate variation (106). The MRP matrix was generated from these trees using CLANN version 2.0.1 (21), and the MRP supertree was generated using PAUP* version 4.0 (98).

Defining ORBs. We define a recombination breakpoint to be a boundary of a genetic region introduced by an LGT event and incorporated via recombination into a genome. If a recombination breakpoint is detected within the boundaries of a gene (here, represented by its annotated open reading frame) using our approach, we refer to that breakpoint as an observable recombination breakpoint (ORB) and classify the corresponding set of homologous genes as ORB^+; these gene sets indicate lateral transfer of a fragment of one or more genes. In contrast, a gene set lacking a detectable internal recombination breakpoint is classified as ORB^-. An ORB^- gene set that has an incongruent phylogeny to a reference species tree indicates lateral transfer of the whole gene and possibly also of genomic sequence extending beyond that gene (see sections below). More detail on the classification of ORB^+ and ORB^- gene sets is given in the study by Chan et al. (15).

Detecting within-gene (fragmentary) genetic transfer. We used a two-phase strategy (17) for detecting recombination within each gene set. Three *P* value statistics—the maximal chi-squared value (71), the neighbor similarity score (52), and the pairwise homoplasy index, as implemented in PhiPack (12)—were first used to detect evidence of recombination events within the sequence sets based on discrepancies in phylogenetic signals. Datasets in which at least two of the three *P* values were ≤ 0.10 were considered as potentially showing evidence of fragmentary LGT. To those gene sets, we applied DualBrothers (72) to define the ORBs more precisely. The program was run with an MCMC 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 1,000. Gene sets 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 ($P \leq 0.05$), 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, $n < m \leq n + 2$; 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

Downloaded from http://jb.asm.org/ on March 30, 2015 by guest

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).

Individual 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 (Mainrole) 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:

$$P(k = x) = f(k;N,m,n) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}$$

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 ρ 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 ρ value of ≈ 0 indicates that the ORB is located at the midpoint of a domain-coding gene region ("domon"; see reference 18), i.e., domon disruption by the ORB, whereas $\rho \approx 1$ 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 ρ 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 $\leq 10^{-100}$ were counted. We use the term "genome 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 (*aroE*), glycerol kinase (*glpF*), guanylate kinase (*gmk*), phosphate acetyltransferase (*pta*), triosephosphate isomerase (*tpi*), and acetyl coenzyme A acetyltransferase (*yqiL*). The seven-gene allelic profile and hence the sequence type (ST) for each *S. aureus* isolate were identified using BLASTN (13) (e-value $\leq 10^{-100}$) against all gene alleles available from http://saureus.mlst.net/ (217 *arcC* alleles, 289 *aroE* alleles, 259 *glpF* 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 supplemental 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.

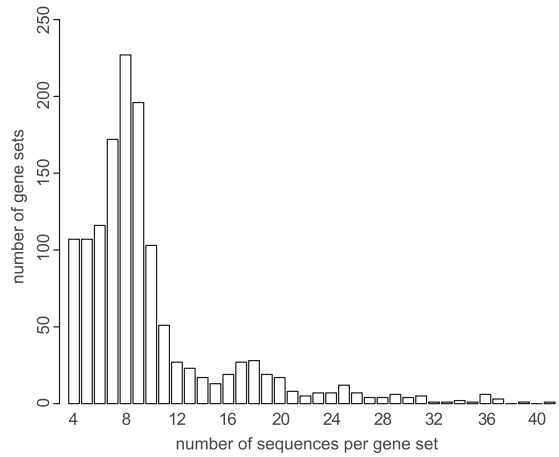

FIG. 1. Size distribution of the 1,354 gene sets examined in the present study.

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; 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

Downloaded from http://jb.asm.org/ on March 30, 2015 by guest

TABLE 2. Inferences based on evidence of within-gene transfer in single-copy and multicopy gene sets[a]

Gene set	Evidence of within-gene transfer	
	Negative	Positive (ORB$^+$ gene set)
SC gene set	No LGT; *SC-noFragGT*	Within-gene LGT between orthologs *SC-FragGT*
MC gene set	No LGT; *MC-noFragGT*	(i) No synologs are recombinant (synolog = paralog) *MC-FragGT-P* (ii) Some synologs are recombinant (synolog = paralog or xenolog) *MC-FragGT-P/X* (iii) All synologs are recombinant (synolog = paralog and/or xenolog) *MC-FragGT-PX*

[a] 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 *-P/X* 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.

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, however, the situation is more complex: (i) if LGT is inferred within the set but topological incongruence with the reference tree is restricted to single-copy genes (i.e., recombination has not affected the synologs), then all synologs are paralogs (*MC-FragGT-P*), and (ii) where both recombinant (xenologous) and nonrecombinant (paralogous) regions of synologs from the same genome are present in a gene set (the recombinant synologs are more precisely paralog-xenolog chimeras), we denote this situation in the set by *MC-FragGT-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 (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* 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 throughout its coding region, although as before 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 for the sequence regions to the immediate left and right are in

TABLE 3. Inference based on evidence of whole-gene transfer in single-copy and multicopy gene sets[a]

Gene set	Evidence of whole-gene transfer	
	Negative	Positive (ORB$^-$ gene set)
SC gene set	No LGT; *SC-noWholeGT*	Whole-gene LGT involving orthologs *SC-WholeGT*
MC gene set	No LGT; *MC-noWholeGT-P*	(i) No recombining sequences are synologs (synolog = paralog) *MC-WholeGT-P* (ii) Some recombining sequences are synologs (synolog = paralog or xenolog) *MC-WholeGT-P/X* (iii) All recombining sequences are synologs (synolog = paralog and/or xenolog) *MC-WholeGT-PX*

[a] 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.

Downloaded from http://jb.asm.org/ on March 30, 2015 by guest

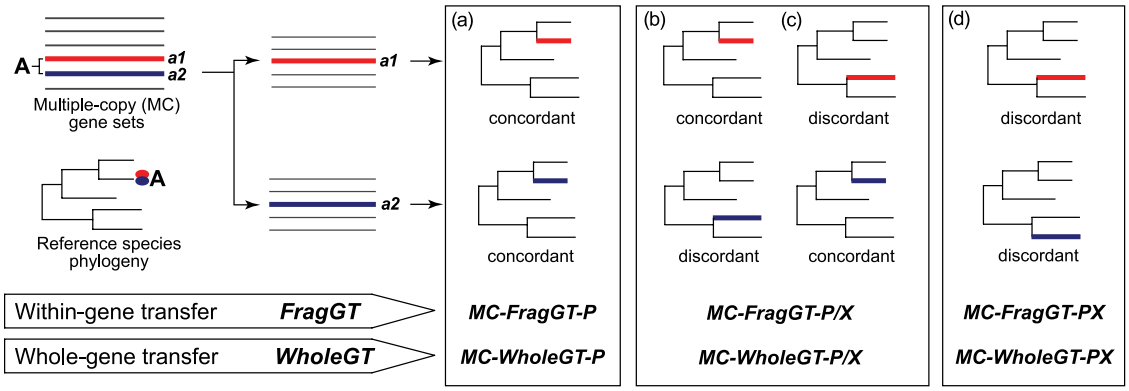

FIG. 2. Testing the source of synologs in multicopy gene sets. The illustrated example is a set with a single synolog, in which genome *A* has two gene copies, *a1* and *a2*. The gene copies are dereplicated in all combinations, yielding two sets of alignments in which each genome is represented only 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 *a1* and *a2* are paralogs; (b and c) one of the trees is concordant, while the other is discordant, suggesting that *a1* and *a2* have different evolutionary trajectories, i.e., *a1* is a paralog, *a2* is a xenolog, or vice versa; and (d) both trees are discordant, suggesting that both *a1* and *a2* 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.)

topological disagreement. These trees can be recovered and compared to the reference supertree, making it straightforward to identify recombinant region(s) within a gene. Where no ORB can be identified, it is necessary to infer trees in a separate step for comparison against the reference topology. As a general rule we carried this out as just described, i.e., by using the synologous gene set as input into the Bayesian inference software MrBayes (89). Two drawbacks of this approach are the resource (CPU and memory) demands associated with inference from large datasets, and the complexity of extracting and interpreting topology data for all minimal subtrees that include synologs, particularly when taking into account the relative support for relevant bipartitions. We realized that there is an opportunity to address both of these drawbacks for the subset of sets (ca. 40% [see below]) that contain only a few (in the present case, one or two) synologs. For each such gene set, we dereplicated synologs in all combinations, yielding a set of alignments in which each genome is represented only once (see Fig. 2 for an example). Then, from each alignment we inferred a Bayesian phylogenetic tree. Since each gene set contains either one or two synologs, for each we generate two to four trees, each of which we compare separately against the reference supertree. In this way we easily automated the extraction of the relevant congruence relationship. This approach, adapted from that of an earlier study (64), could in principle be extended to greater numbers of synologs per set, although at exponential cost unless additional filtering steps are implemented.

Within-gene (fragmentary) genetic transfer. We applied a two-phase strategy (17) to detect recombination in each of the 1,354 gene sets, the first phase involving three statistical tests for inferring phylogenetic discrepancies and the second involving a Bayesian approach to more-accurately identify ORBs (see Materials and Methods). Of the 1,354 gene sets, we found 401 (29.6%) that show evidence of within-gene recombination (i.e., sets containing one or more ORBs) after first-phase screening for phylogenetic discrepancies. Of these sets, 252 (18.6% of 1,354) also showed clear evidence of recombination based on Bayesian phylogenetic analysis in the second phase, with Bayesian posterior probability (BPP) support of ≥0.500 for the dominant topology (as determined internally with respect to the individual alignment) on at least one side of the inferred breakpoint in each set. These are the ORB^+ gene sets. For 68 sets (5.0%), we could identify an ORB, but no sequence region has a BPP of ≥0.500; we labeled these as inconclusive. For a further 81 gene sets (6.0%), recombination was detected in the first phase, but no ORB could be identified in the second phase (i.e., these are the false positives from the first phase). No evidence of recombination was detected in 953 (70.4%) gene sets in first-phase screening.

Of the 252 ORB^+ gene sets that show clear evidence of within-gene (fragmentary) genetic transfer, 98 are single-copy gene sets (*SC-FragGT*), while the other 154 are multicopy gene sets (*MC-FragGT*). Phylogenetic discrepancy at different regions across a gene set alignment can sometimes be due to regional (e.g., domain-specific) differences in rates of nucleotide substitution in one or more sequences. We looked specifically for such differential rates ($\mu \geq 0.30$, from DualBrothers) but did not observe any instances that span the inferred breakpoints. This suggests that the breakpoints inferred here indeed arise from genetic recombination and are not artifacts of variation in the rate of nucleotide substitution.

To examine possible functional bias pertaining to fragmentary genetic transfer within the 252 sets, we used annotations from the JCVI Comprehensive Microbial Resource (http://cmr.jcvi.org/) to assign a functional category (Mainrole) to each gene set. Figure 3 shows the proportions of proteins in each functional category for (Fig. 3a) single-copy (*SC-FragGT*) and (Fig. 3b) multicopy (*MC-FragGT*) gene sets for which we inferred fragmentary genetic transfer, compared to their frequencies in the full (1,354-set) staphylococcal data set.

Gene sets affected by fragmentary genetic transfer are significantly either over- or under-represented in more than half of the JCVI functional categories, and this is the case for both

Downloaded from http://jb.asm.org/ on March 30, 2015 by guest

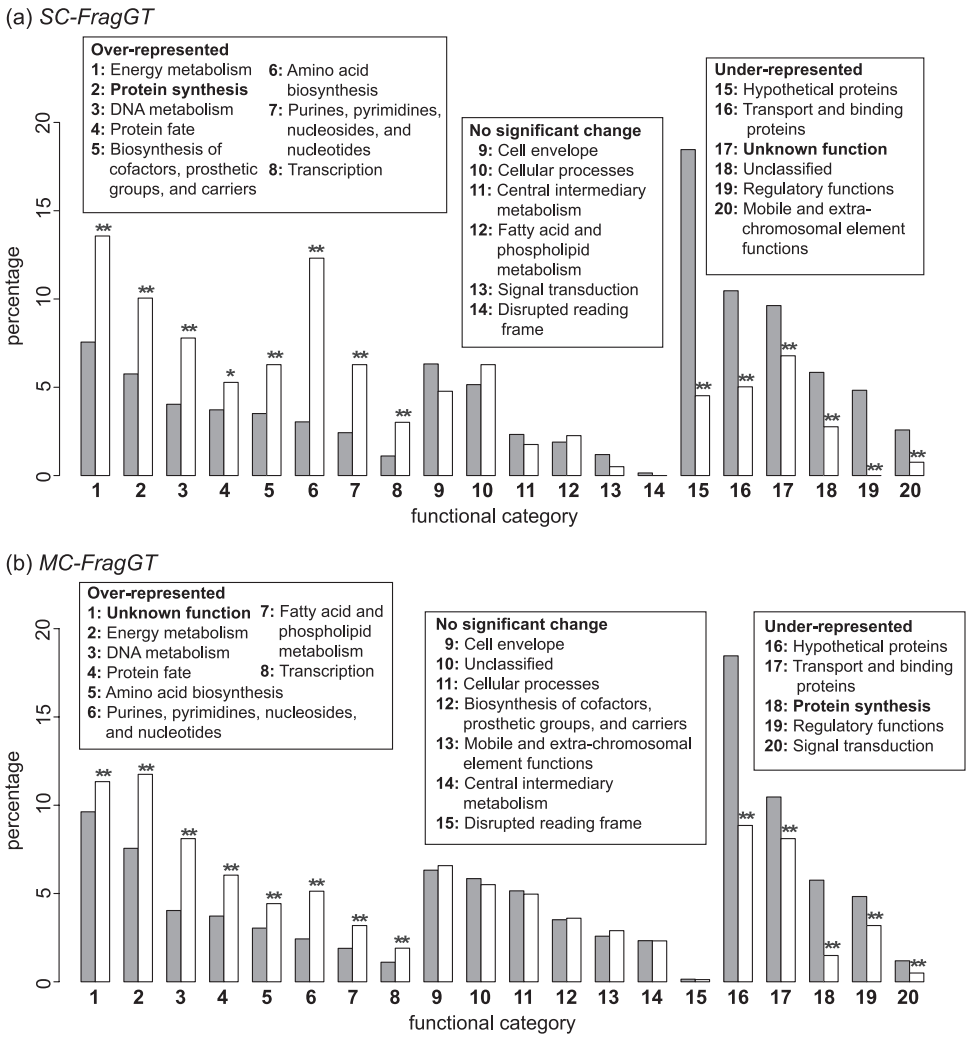

FIG. 3. Representation of functional categories assigned to protein sequences corresponding to gene sets in *Staphylococcus* species that show evidence of within-gene (fragmentary) genetic transfer (open bars) for single-copy gene sets (*SC-FragGT*) (a) and multicopy gene sets (*MC-FragGT*) (b). The solid bars show these same functional categories in the full data set (1,354 sets, 13,297 proteins). Categories are numbered differently for panels a and b as 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.

SC and MC gene sets. 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

Downloaded from http://jb.asm.org/ on March 30, 2015 by guest

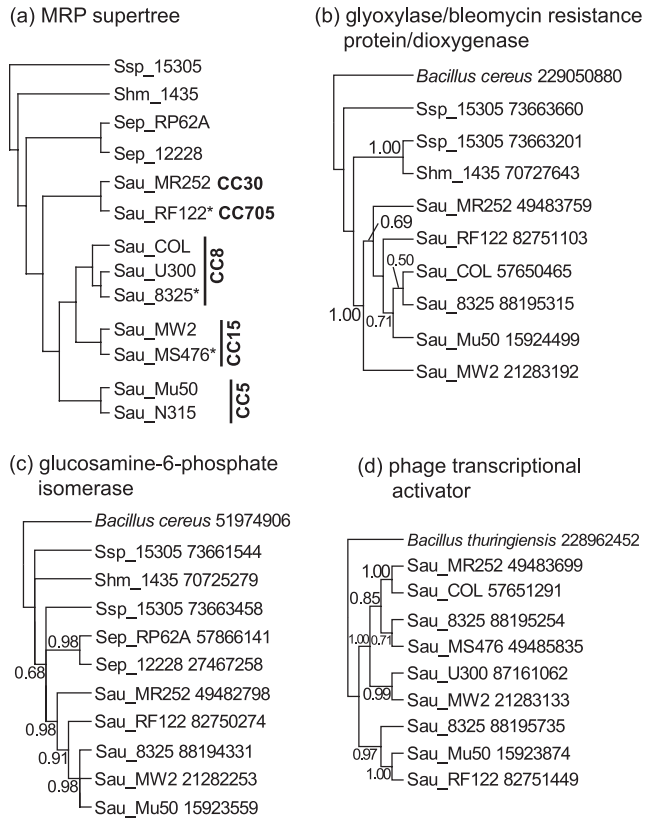

FIG. 4. Reference tree used in the present study, and three instances of tree topologies showing history of whole-gene transfer: the supertree for the 13 staphylococcal genomes (methicillin-sensitive *S. aureus* isolates are marked with an asterisk), rooted based on a previous phylogenetic study of *Staphylococcus* species using small subunit rRNA genes (99) that gives *S. saprophyticus* as the outgroup (a), and tree topologies for gene sets encoding for glyoxalase/bleomycin resistance protein/dioxygenases (gene set 445), an instance of *MC-WholeGT-P* (the synolog is a paralog) (b); glucosamine-6-phosphate isomerase (gene set 1167), an instance of *MC-WholeGT-P/X* (the synolog is either a paralog or a xenolog, but not both) (c); and phage transcriptional activator (gene set 267), an instance of *MC-WholeGT-PX* (the synolog is a paralog and/or a xenolog) (d). GenBank GI numbers are shown for all sequences implicated in each of the phylogenies b through d, in which a sequence from *Bacillus cereus* or *B. thuringiensis* is used as an outgroup. Labels of different genome isolates follow the description in Table 1. Bayesian posterior probability values of ≥0.50 are shown at the internal nodes.

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 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-P/X*), 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-P/X*), in which a copy shows evidence of gene sharing with lineages of *S. epidermidis* or *S. aureus* (i.e., is a xenolog), whereas another

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[a]

Category	Tree comparison based on maximum-likelihood tests	
	Concordant	Discordant
SC gene sets, no synologs	*SC-noWholeGT*, 677 sets	*SC-WholeGT*, 97 sets
Gene sets with one or two synologs each (within MC gene sets)	*MC-noWholeGT-P*, 47 sets	*MC-WholeGT*, 58 sets
Gene sets with more than two synologs each (within MC gene sets)	*MC-noWholeGT-P*, 3 sets	*MC-WholeGT*, 152 sets
Total no. of gene sets (%)	727 (53.7)	307 (22.7)

[a] 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.

Downloaded from http://jb.asm.org/ on March 30, 2015 by guest

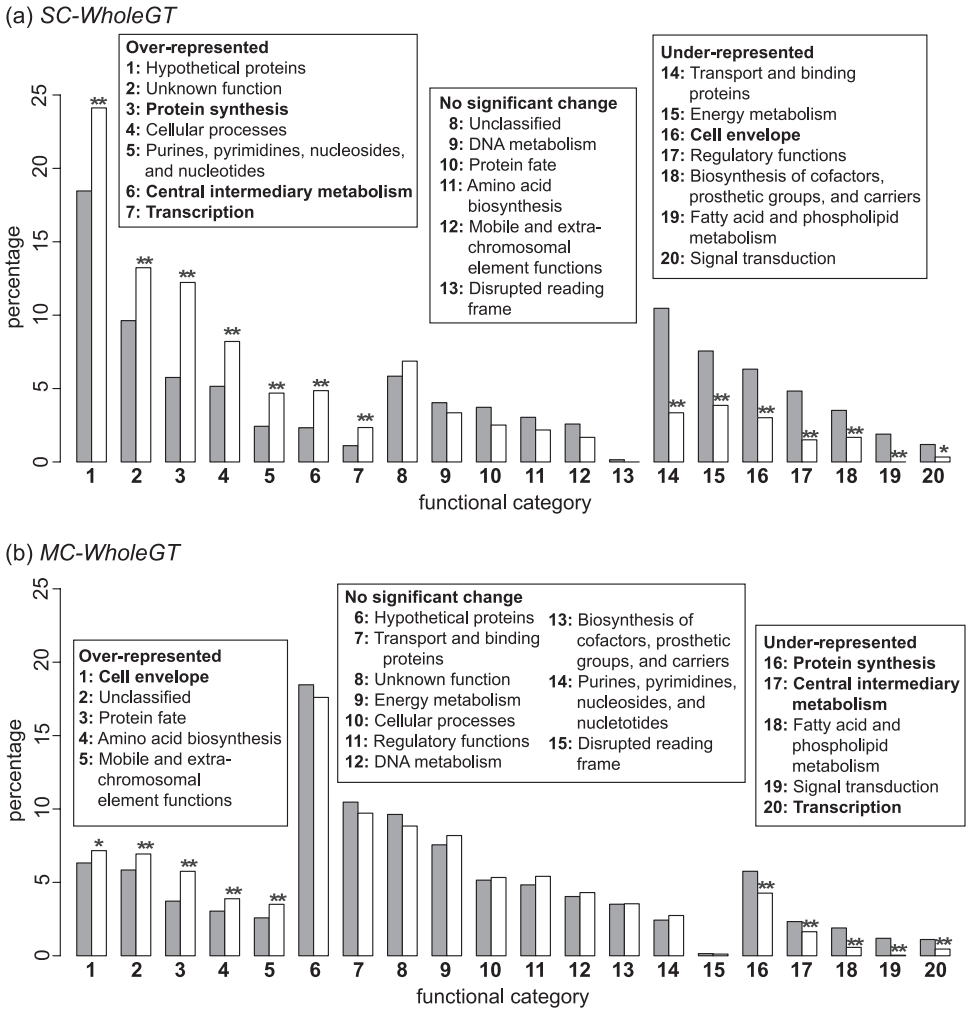

FIG. 5. Representation of functional categories assigned to protein sequences corresponding to gene sets in *Staphylococcus* species that show evidence of whole-gene transfer (□) for single-copy gene sets (*SC-WholeGT*) (a) and multicopy gene sets (*MC-WholeGT*) (b). The solid bars (▨) show the same 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.

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 *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 metabolism, and transcription categories (in contrast to under-representation of these same categories in MC sets), suggesting that *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 functional categories are significantly under-represented in *MC-WholeGT*, suggesting that these genomes have been less receptive 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, including energy metabolism, are under-represented. The categories cell envelope and mobile and extrachromosomal genetic element functions are over-represented in *MC-WholeGT*. The

Downloaded from http://jb.asm.org/ on March 30, 2015 by guest

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), especially cation transport, e.g., subunits of the monovalent cation/proton antiporter proteins.

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,

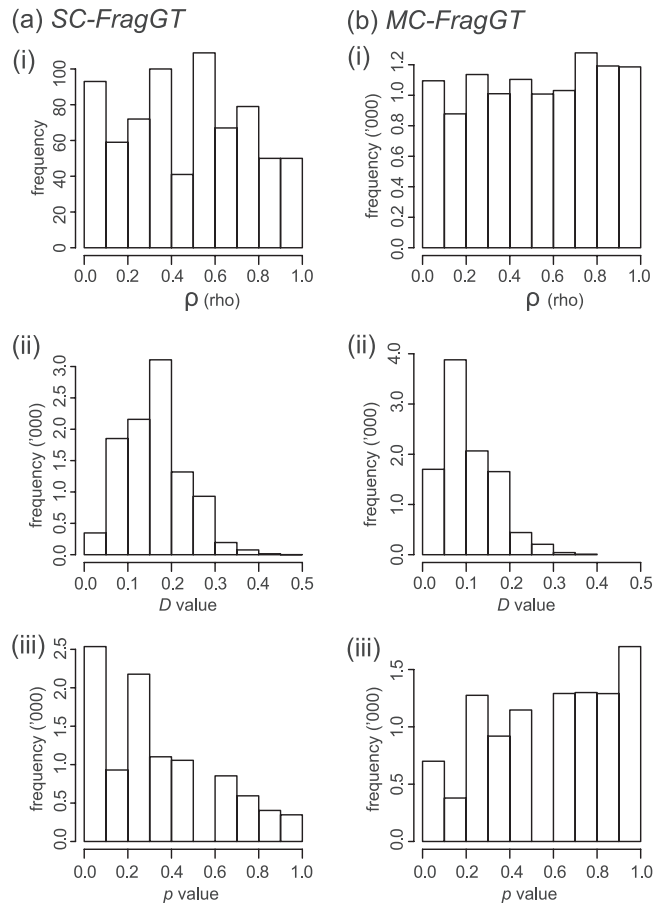

FIG. 6. Tendency of domon disruption by LGT as observed in single-copy gene sets (*SC-FragGT*) (a) and multicopy gene sets (*MC-FragGT*) (b) in *Staphylococcus* genomes. Shown, respectively, for each in panels a and b are the observed ρ values (large ρ values indicate little domon disruption) (i) and the distributions of *D* (ii) and *P* (iii) values generated via 10,000 Kolmogorov-Smirnov tests to examine statistical differences between the distributions of observed (randomly subsampled) and expected (uniformly distributed) ρ values.

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 ($P \leq 0.01$) over-represented in *MC-WholeGT* (Fig. 5b) and most MGE genes show high suscepti-

Downloaded from http://jb.asm.org/ on March 30, 2015 by guest

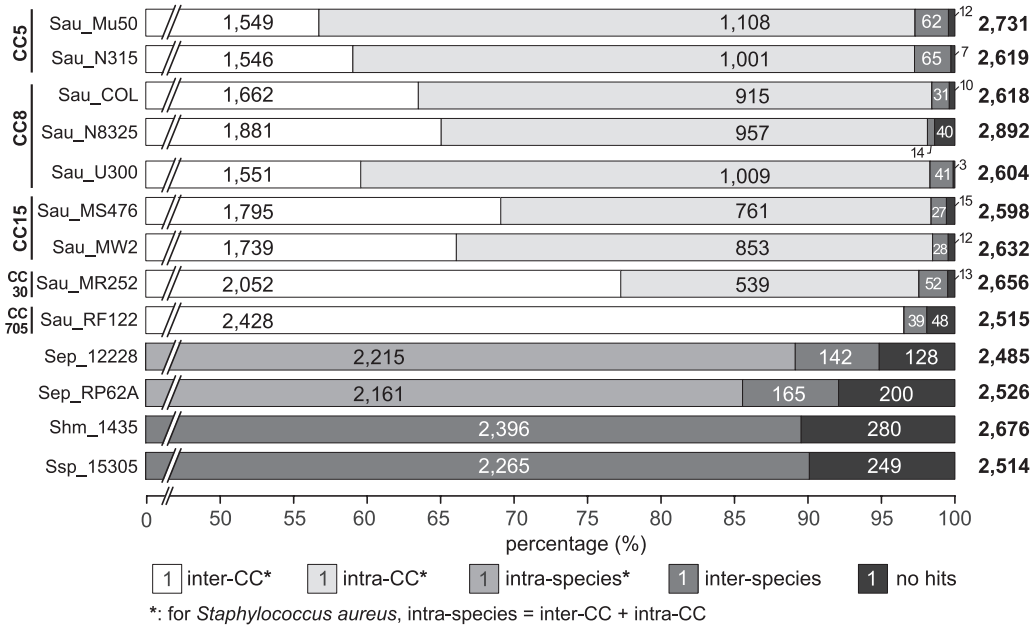

FIG. 7. Genome affinities of the 13 staphylococcal genomes based on analysis of sequence similarity. The bar for each of the 13 genomes depicts the percentage of proteins (numbers are shown in the bar) showing high similarity to sequences in other clonal complexes (CCs; inter-CC), within the same CC (intra-CC), within the same species (intraspecies), and in other species (interspecies), as well as showing no matches to other staphylococcal isolates or species (see Table S5 in the supplemental material). Labels of different genome isolates follow the description in Table 1. Instances of inter-CC and intra-CC genome affinity apply only to *S. aureus*. In these cases (the top nine bars representing the *S. aureus* genomes), intraspecies genome affinities equal the sum of cases for inter-CC and intra-CC.

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 *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 *S. aureus* and *S. epidermidis* find a best match or matches within the same species, suggesting high level of intraspecies genome affinity. *S. haemolyticus* and *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 *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 *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 *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 *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 *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

Downloaded from http://jb.asm.org/ on March 30, 2015 by guest

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 (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 LGT 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 [*glpF*], 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

Downloaded from http://jb.asm.org/ on March 30, 2015 by guest

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 *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 *Staphylococcus*, including virulently pathogenic isolates.

ACKNOWLEDGMENTS

This study was supported by Australian Research Council grant CE0348221.

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

REFERENCES

1. **Andreeva, A., et al.** 2004. SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res. **32:**D226–D229.
2. **Araújo, S. A., et al.** 2010. Fatal staphylococcal infection following classic dengue fever. Am. J. Trop. Med. Hyg. **83:**679–682.
3. **Archibald, J. M., and A. J. Roger.** 2002. Gene duplication and gene conversion shape the evolution of archaeal chaperonins. J. Mol. Biol. **316:** 1041–1050.
4. **Baba, T., et al.** 2002. Genome and virulence determinants of high virulence community-acquired MRSA. Lancet **359:**1819–1827.
5. **Bapteste, E., et al.** 2009. Prokaryotic evolution and the tree of life are two different things. Biol. Direct **4:**34.
6. **Barlow, M.** 2009. What antimicrobial resistance has taught us about horizontal gene transfer. Methods Mol. Biol. **532:**397–411.
7. **Beiko, R. G., C. X. Chan, and M. A. Ragan.** 2005. A word-oriented approach to alignment validation. Bioinformatics **21:**2230–2239.
8. **Beiko, R. G., T. J. Harlow, and M. A. Ragan.** 2005. Highways of gene sharing in prokaryotes. Proc. Natl. Acad. Sci. U. S. A. **102:**14332–14337.
9. **Bloemendaal, A. L. A., E. C. Brouwer, and A. C. Fluit.** 2010. Methicillin resistance transfer from *Staphylocccus epidermidis* to methicillin-susceptible *Staphylococcus aureus* in a patient during antibiotic therapy. PLoS One **5:**e11841.
10. **Bork, P., and R. F. Doolittle.** 1992. Proposed acquisition of an animal protein domain by bacteria. Proc. Natl. Acad. Sci. U. S. A. **89:**8990–8994.
11. **Brody, T., et al.** 2008. Horizontal gene transfers link a human MRSA pathogen to contagious bovine mastitis bacteria. PLoS One **3:**e3074.
12. **Bruen, T. C., H. Philippe, and D. Bryant.** 2006. A simple and robust statistical test for detecting the presence of recombination. Genetics **172:** 2665–2681.
13. **Camacho, C., et al.** 2009. BLAST+: architecture and applications. BMC Bioinform. **10:**421.
14. **Chambers, H. F., and F. R. Deleo.** 2009. Waves of resistance: *Staphylococcus aureus* in the antibiotic era. Nat. Rev. Microbiol. **7:**629–641.
15. **Chan, C. X., R. G. Beiko, A. E. Darling, and M. A. Ragan.** 2009. Lateral transfer of genes and gene fragments in prokaryotes. Genome Biol. Evol. **1:**429–438.
16. **Chan, C. X., R. G. Beiko, and M. A. Ragan.** 2006. Detecting recombination in evolving nucleotide sequences. BMC Bioinform. **7:**412.
17. **Chan, C. X., R. G. Beiko, and M. A. Ragan.** 2007. A two-phase strategy for detecting recombination in nucleotide sequences. S. Afr. Comput. J. **38:**20–27.
18. **Chan, C. X., A. E. Darling, R. G. Beiko, and M. A. Ragan.** 2009. Are protein domains modules of lateral genetic transfer? PLoS One **4:**e4524.
19. **Chen, J., and R. P. Novick.** 2009. Phage-mediated intergeneric transfer of toxin genes. Science **323:**139–141.
20. **Cooper, J. E., and E. J. Feil.** 2006. The phylogeny of *Staphylococcus aureus*: which genes make the best intra-species markers? Microbiology **152:**1297–1305.
21. **Creevey, C. J., et al.** 2004. Does a tree-like phylogeny only exist at the tips in the prokaryotes? Proc. R. Soc. Lond. B Biol. Sci. **271:**2551–2558.
22. **Davies, J., and D. Davies.** 2010. Origins and evolution of antibiotic resistance. Microbiol. Mol. Biol. Rev. **74:**417–433.
23. **Deresinski, S.** 2009. Vancomycin heteroresistance and methicillin-resistant *Staphylococcus aureus.* J. Infect. Dis. **199:**605–609.
24. **Diep, B. A., et al.** 2006. Complete genome sequence of USA300, an epidemic clone of community-acquired meticillin-resistant *Staphylococcus aureus*. Lancet **367:**731–739.
25. **Donegan, N. P., and A. L. Cheung.** 2009. Regulation of the *mazEF* toxin-antitoxin module in *Staphylococcus aureus* and its impact on *sigB* expression. J. Bacteriol. **191:**2795–2805.
26. **Edgar, R. C.** 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. **32:**1792–1797.
27. **Enright, A. J., S. Van Dongen, and C. A. Ouzounis.** 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. **30:**1575–1584.
28. **Feil, E. J., et al.** 2003. How clonal is *Staphylococcus aureus*? J. Bacteriol. **185:**3307–3316.
29. **Feil, E. J., B. C. Li, D. M. Aanensen, W. P. Hanage, and B. G. Spratt.** 2004. eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. J. Bacteriol. **186:**1518–1530.
30. **Feil, E. J., and B. G. Spratt.** 2001. Recombination and the population structures of bacterial pathogens. Annu. Rev. Microbiol. **55:**561–590.
31. **Feng, Y., et al.** 2008. Evolution and pathogenesis of *Staphylococcus aureus*: lessons learned from genotyping and comparative genomics. FEMS Microbiol. Rev. **32:**23–37.
32. **Finn, R. D., et al.** 2006. Pfam: clans, web tools, and services. Nucleic Acids Res. **34:**D247–D251.
33. **Fitch, W. M.** 2000. Homology: a personal view on some of the problems. Trends Genet. **16:**227–231.
34. **Fitzpatrick, D. A.** 2009. Lines of evidence for horizontal gene transfer of a phenazine producing operon into multiple bacterial species. J. Mol. Evol. **68:**171–185.
35. **Fu, Z., S. Tamber, G. Memmi, N. P. Donegan, and A. L. Cheung.** 2009. Overexpression of MazF in *Staphylococcus aureus* induces bacteriostasis by selectively targeting mRNAs for cleavage. J. Bacteriol. **191:**2051–2059.
36. **Gill, S. R., et al.** 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.
37. **Goldman, N., J. P. Anderson, and A. G. Rodrigo.** 2000. Likelihood-based tests of topologies in phylogenetics. Syst. Biol. **49:**652–670.
38. **Götz, S., et al.** 2008. High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res. **36:**3420–3435.
39. **Gray, G. S., and W. M. Fitch.** 1983. Evolution of antibiotic resistance genes: the DNA sequence of a kanamycin resistance gene from *Staphylococcus aureus*. Mol. Biol. Evol. **1:**57–66.
40. **Harlow, T. J., J. P. Gogarten, and M. A. Ragan.** 2004. A hybrid clustering approach to recognition of protein families in 114 microbial genomes. BMC Bioinform. **5:**45.
41. **Hartl, D. L., E. R. Lozovskaya, and J. G. Lawrence.** 1992. Nonautonomous transposable elements in prokaryotes and eukaryotes. Genetica **86:**47–53.
42. **Herron, L. L., et al.** 2002. Genome sequence survey identifies unique sequences and key virulence genes with unusual rates of amino acid substitution in bovine *Staphylococcus aureus*. Infect. Immun. **70:**3978–3981.
43. **Higgins, C. F.** 1992. ABC transporters: from microorganisms to man. Annu. Rev. Cell Biol. **8:**67–113.
44. **Hiramatsu, K., S. Watanabe, F. Takeuchi, T. Ito, and T. Baba.** 2004. Genetic characterization of methicillin-resistant *Staphylococcus aureus*. Vaccine **22:**S5–S8.
45. **Holden, M. T., et al.** 2004. Complete genomes of two clinical *Staphylococcus aureus* strains: evidence for the rapid evolution of virulence and drug resistance. Proc. Natl. Acad. Sci. U. S. A. **101:**9786–9791.
46. **Hooper, S. D., and O. G. Berg.** 2003. Duplication is more common among laterally transferred genes than among indigenous genes. Genome Biol. **4:**R48.
47. **Iandolo, J. J., et al.** 2002. Comparative analysis of the genomes of the temperate bacteriophages phi 11, phi 12, and phi 13 of *Staphylococcus aureus* 8325. Gene **289:**109–118.
48. **Igarashi, N., et al.** 2001. Horizontal transfer of the photosynthesis gene cluster and operon rearrangement in purple bacteria. J. Mol. Evol. **52:**333–341.
49. **Inagaki, Y., E. Susko, and A. J. Roger.** 2006. Recombination between elongation factor 1-alpha genes from distantly related archaeal lineages. Proc. Natl. Acad. Sci. U. S. A. **103:**4528–4533.
50. **Ito, T., Y. Katayama, and K. Hiramatsu.** 1999. Cloning and nucleotide sequence determination of the entire *mec* DNA of pre-methicillin-resistant *Staphylococcus aureus* N315. Antimicrob. Agents Chemother. **43:**1449–1458.
51. **Jain, R., M. C. Rivera, and J. A. Lake.** 1999. Horizontal gene transfer among genomes: the complexity hypothesis. Proc. Natl. Acad. Sci. U. S. A. **96:**3801–3806.
52. **Jakobsen, I. B., and S. Easteal.** 1996. A program for calculating and dis-

Downloaded from http://jb.asm.org/ on March 30, 2015 by guest

playing compatibility matrices as an aid in determining reticulate evolution in molecular sequences. Comput. Appl. Biosci. **12:**291–295.
53. **Johnson, N. L., S. Kotz, and A. W. Kemp.** 1992. Univariate discrete distributions, 2nd ed. Wiley, New York, NY.
54. **Kamada, K., F. Hanaoka, and S. K. Burley.** 2003. Crystal structure of the MazE/MazF complex: molecular bases of antidote-toxin recognition. Mol. Cell **11:**875–884.
55. **Kern, W. V.** 2010. Management of *Staphylococcus aureus* bacteremia and endocarditis: progresses and challenges. Curr. Opin. Infect. Dis. **23:**346–358.
56. **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.
57. **Kishino, H., and M. Hasegawa.** 1989. Evaluation of the maximum-likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in *Hominoidea*. J. Mol. Evol. **29:**170–179.
58. **Koonin, E. V.** 2005. Orthologs, paralogs, and evolutionary genomics. Annu. Rev. Genet. **39:**309–338.
59. **Koonin, E. V., and Y. I. Wolf.** 2009. The fundamental units, processes and patterns of evolution, and the tree of life conundrum. Biol. Direct **4:**33.
60. **Kuroda, M., et al.** 2001. Whole genome sequencing of meticillin-resistant *Staphylococcus aureus*. Lancet **357:**1225–1240.
61. **Kuroda, M., et al.** 2005. Whole genome sequence of *Staphylococcus saprophyticus* reveals the pathogenesis of uncomplicated urinary tract infection. Proc. Natl. Acad. Sci. U. S. A. **102:**13272–13277.
62. **Lacey, R. W.** 1980. Evidence for two mechanisms of plasmid transfer in mixed cultures of *Staphylococcus aureus*. J. Gen. Microbiol. **119:**423–435.
63. **Lefébure, T., P. D. Bitar, H. Suzuki, and M. J. Stanhope.** 2010. Evolutionary dynamics of complete *Campylobacter* pan-genomes and the bacterial species concept. Genome Biol. Evol. **2:**646–655.
64. **Lerat, E., V. Daubin, H. Ochman, and N. A. Moran.** 2005. Evolutionary origins of genomic repertoires in bacteria. PLoS Biol. **3:**e130.
65. **Lindsay, J. A.** 2010. Genomic variation and evolution of *Staphylococcus aureus*. Int. J. Med. Microbiol. **300:**98–103.
66. **Lindsay, J. A., and M. T. G. Holden.** 2004. *Staphylococcus aureus*: superbug, super genome? Trends Microbiol. **12:**378–385.
67. **Lowy, F. D.** 1998. *Staphylococcus aureus* infections. N. Engl. J. Med. **339:** 520–532.
68. **Lynch, M., and J. S. Conery.** 2000. The evolutionary fate and consequences of duplicate genes. Science **290:**1151–1155.
69. **Malachowa, N., and F. R. DeLeo.** 2010. Mobile genetic elements of *Staphylococcus aureus*. Cell. Mol. Life Sci. **67:**3057–3071.
70. **Massey, F. J.** 1951. The Kolmogorov-Smirnov test for goodness of fit. J. Am. Stat. Assoc. **46:**68–78.
71. **Maynard Smith, J.** 1992. Analyzing the mosaic structure of genes. J. Mol. Evol. **34:**126–129.
72. **Minin, V. N., K. S. Dorman, F. Fang, and M. A. Suchard.** 2005. Dual multiple change-point model leads to more accurate recombination detection. Bioinformatics **21:**3034–3042.
73. **Morris, R. T., and G. Drouin.** 2007. Ectopic gene conversions in bacterial genomes. Genome **50:**975–984.
74. **Niazi, S. A., et al.** 2010. *Propionibacterium acnes* and *Staphylococcus epidermidis* isolated from refractory endodontic lesions are opportunistic pathogens. J. Clin. Microbiol. **48:**3859–3869.
75. **Nikoh, N., and A. Nakabachi.** 2009. Aphids acquired symbiotic genes via lateral gene transfer. BMC Biol. **7:**12.
76. **Nilsson, A. I., et al.** 2006. Reducing the fitness cost of antibiotic resistance by amplification of initiator tRNA genes. Proc. Natl. Acad. Sci. U. S. A. **103:**6976–6981.
77. **Noto, M. J., P. M. Fox, and G. L. Archer.** 2008. Spontaneous deletion of the methicillin resistance determinant, *mecA*, partially compensates for the fitness cost associated with high-level vancomycin resistance in *Staphylococcus aureus*. Antimicrob. Agents Chemother. **52:**1221–1229.
78. **Notredame, C., D. G. Higgins, and J. Heringa.** 2000. T-Coffee: a novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. **302:**205–217.
79. **Novick, R. P., G. E. Christie, and J. R. Penades.** 2010. The phage-related chromosomal islands of Gram-positive bacteria. Nat. Rev. Microbiol. **8:**541–551.
80. **Novick, R. P., and A. Subedi.** 2007. The SaPIs: mobile pathogenicity islands of *Staphylococcus*. Chem. Immunol. Allergy **93:**42–57.
81. **Omelchenko, M. V., K. S. Makarova, Y. I. Wolf, I. B. Rogozin, and E. V. Koonin.** 2003. Evolution of mosaic operons by horizontal gene transfer and gene displacement *in situ*. Genome Biol. **4:**55.
82. **Otto, M.** 2009. *Staphylococcus epidermidis*: the "accidental" pathogen. Nat. Rev. Microbiol. **7:**555–567.
83. **Palmer, K. L., V. N. Kos, and M. S. Gilmore.** 2010. Horizontal gene transfer and the genomics of enterococcal antibiotic resistance. Curr. Opin. Microbiol. **13:**632–639.
84. **Périchon, B., and P. Courvalin.** 2009. VanA-type vancomycin-resistant *Staphylococcus aureus*. Antimicrob. Agents Chemother. **53:**4580–4587.
85. **Puigbo, P., Y. I. Wolf, and E. V. Koonin.** 2010. The tree and net components of prokaryote evolution. Genome Biol. Evol. **2:**745–756.
86. **Queck, S. Y., et al.** 2009. Mobile genetic element-encoded cytolysin connects virulence to methicillin resistance in MRSA. PLoS Pathog. **5:**e1000533.
87. **Ragan, M. A.** 1992. Phylogenetic inference based on matrix representation of trees. Mol. Phylogenet. Evol. **1:**53–58.
88. **Ragan, M. A., and R. G. Beiko.** 2009. Lateral genetic transfer: open issues. Philos. Trans. R. Soc. Lond. B Biol. Sci. **364:**2241–2251.
89. **Ronquist, F., and J. P. Huelsenbeck.** 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics **19:**1572–1574.
90. **Saier, M. H., Jr., and J. Reizer.** 1994. The bacterial phosphotransferase system: new frontiers 30 years later. Mol. Microbiol. **13:**755–764.
91. **Sandegren, L., and D. I. Andersson.** 2009. Bacterial gene amplification: implications for the evolution of antibiotic resistance. Nat. Rev. Microbiol. **7:**578–588.
92. **Schmidt, H. A., K. Strimmer, M. Vingron, and A. von Haeseler.** 2002. TREE-PUZZLE: maximum-likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics **18:**502–504.
93. **Shimodaira, H., and M. Hasegawa.** 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol. Biol. Evol. **16:**1114–1116.
94. **Shinefield, H. R., and N. L. Ruff.** 2009. Staphylococcal infections: a historical perspective. Infect. Dis. Clin. N. Am. **23:**1–15.
95. **Skippington, E., and M. A. Ragan.** Lateral genetic transfer and the construction of genetic exchange communities. FEMS Microbiol. Rev., in press. doi:10.1111/j.1574-6976.2010.00261.x.
96. **Stout, V. G., and J. J. Iandolo.** 1990. Chromosomal gene transfer during conjugation by *Staphylococcus aureus* is mediated by transposon-facilitated mobilization. J. Bacteriol. **172:**6148–6150.
97. **Strimmer, K., and A. Rambaut.** 2002. Inferring confidence sets of possibly misspecified gene trees. Proc. R. Soc. Lond. B Biol. Sci. **269:**137–142.
98. **Swofford, D. L.** 2003. PAUP*: phylogenetic analysis using parsimony (*and other methods), 4th ed. Sinauer Associates, Sunderland, MA.
99. **Takahashi, T., I. Satoh, and N. Kikuchi.** 1999. Phylogenetic relationships of 38 taxa of the genus *Staphylococcus* based on 16S rRNA gene sequence analysis. Int. J. Syst. Bacteriol. **49:**725–728.
100. **Takeuchi, F., et al.** 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.
101. **Van Melderen, L., and M. Saavedra De Bast.** 2009. Bacterial toxin-antitoxin systems: more than selfish entities? PLoS Genet. **5:**e1000437.
102. **Waness, A.** 2010. Revisiting methicillin-resistant *Staphylococcus aureus* infections. J. Glob. Infect. Dis. **2:**49–56.
103. **Woese, C. R.** 2000. Interpreting the universal phylogenetic tree. Proc. Natl. Acad. Sci. U. S. A. **97:**8392–8396.
104. **Woolfit, M., I. Iturbe-Ormaetxe, E. A. McGraw, and S. L. O'Neill.** 2009. An ancient horizontal gene transfer between mosquito and the endosymbiotic bacterium *Wolbachia pipientis*. Mol. Biol. Evol. **26:**367–374.
105. **Wright, G. D.** 2010. Antibiotic resistance in the environment: a link to the clinic? Curr. Opin. Microbiol. **13:**589–594.
106. **Yang, Z.** 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. **39:**306–314.
107. **Zhang, Y. Q., et al.** 2003. Genome-based analysis of virulence genes in a non-biofilm-forming *Staphylococcus epidermidis* strain (ATCC 12228). Mol. Microbiol. **49:**1577–1593.
108. **Zhaxybayeva, O., J. P. Gogarten, R. L. Charlebois, W. F. Doolittle, and R. T. Papke.** 2006. Phylogenetic analyses of cyanobacterial genomes: quantification of horizontal gene transfer events. Genome Res. **16:**1099–1108.

Downloaded from http://jb.asm.org/ on March 30, 2015 by guest