Appendix

Table of Contents

Supporting Tables

Table S.1. Descriptions of all 35 point mutations ordered by time of appearance in the JH isolates. See notes beneath Table.

No.*	Type of mutation†	Mutated locus‡	Description of mutated locus	Mutation(s)
(a) Appeared first in JH1				
(i) Loci involved in β-lactam resistance				
1	FRAME	SAP011 (*blaR1*) (on plasmid)	involved in regulation of β-lactamase gene *blaZ* and broad spectrum β-lactam resistance gene *mecA* (*1-3*)	A nucleotide deletion of an A at the 466th nucleotide position in the gene frameshifted the last 70% of the gene. The deletion occurred in a homopolymeric tract TAAAAAAAAT.
(b) Appeared first in JH2				
(i) Loci previously implicated in both vancomycin and β-lactam resistance				
2	NONSYN	SA1702	in operon with gene *vraR*, which is possibly involved in regulation of cell wall synthesis (*4-7*)	amino acid change H164R caused by a nucleotide substitution A→G at the 491st nucleotide position in the gene
(ii) Loci previously implicated in rifampin resistance				
3	NONSYN	SA0500 (*rpoB*)	codes for β-subunit of RNA polymerase (*8-15*)	amino acid change D471Y caused by a nucleotide substitution G→T at the 1411th nucleotide position in the gene
4	NONSYN			amino acid change A473S caused by a nucleotide substitution G→T at the 1417th nucleotide position in the gene
5	NONSYN			amino acid change A477S caused by a nucleotide substitution G→T at the 1429th nucleotide position in the gene
6	NONSYN			amino acid change E478D caused by a nucleotide substitution G→T at the 1434th nucleotide position in the gene
(iii) Loci previously implicated in daptomycin resistance				
7	NONSYN	SA0501 (*rpoC*)	codes for β'-subunit of RNA polymerase (*11, 16*)	amino acid change E854K caused by a nucleotide substitution G→A at the 2560th nucleotide position in the gene, may map to outer rim of secondary channel (*16*)
(iv) Other loci				
8	NONSYN	SA1129	contains match to RNA binding motif	amino acid change D296Y caused by a nucleotide substitution G→T at the 886th nucleotide position in the gene
(c) Appeared first in JH5				
9	FRAME	SA1249	putative gene possibly in an operon with the gene *murG*, which is involved in cell wall synthesis	A nucleotide deletion of a G at the 31st nucleotide position in the gene frameshifted the last 80% of the gene. The deletion occurred in a homopolymeric tract TGGGGGGGGGA.
(d) Appeared first in JH6				
(i) Loci previously implicated in vancomycin resistance				
10	FRAME	SA1843 (*agrC*)	in the *agr* locus involved in qorum sensing and regulation of the expression of virulence and cell surface proteins (*17-25*)	A nucleotide deletion of a T at the 313th nucleotide position in the gene frameshifted the last 70% of the gene. The deletion occurred in a homopolymeric tract ATTTTTTTA.
(ii) Loci previously implicated in daptomycin resistance				
11	TRUNC	SA0019 (*yycH*)	in a gene cluster with the gene *yycF* and possibly involved in the regulation of the autolysin gene *lytM* (*11, 26-29*)	A nucleotide substitution G→A at the 107th nucleotide position in the gene converted the 36th codon to a stop codon, possibly truncating protein to 10% of its length.
(iii) Other loci				
12	INT DIV	between divergently transcribed genes SAS014 and SA0411 (*ndhF*)	NdhF is the F-subunit of NADH dehydrogenase.	nucleotide deletion of a T 579- and 452-bp upstream of SAS014 and SA0411 respectively, occurred in a homopolymeric tract ATTTTTTTC
13	SYN	SA0582	similar to Na^+/H^+ antiporter subunit MrpE in *B. subtilis* (*30*)	nucleotide substitution T→C at the 90th nucleotide position in the

				30th codon
14	NONSYN	SA0980 (*isdE*)	involved in passage of heme-iron to cytoplasm during pathogenesis (*31-33*)	amino acid change A84V caused by a nucleotide substitution C→T at the 251st nucleotide position in the gene
15	FRAME	SA1659 (*prsA*)	codes for a chaperone that assists post-translocational folding of proteins at the cytoplasmic/cell wall interface (*34-39*)	A nucleotide deletion of an A at the 804th nucleotide position in the gene frameshifted the last 15% of the gene. The deletion occurred in a homopolymeric tract CAAAAAAAT.
16	NONSYN	SA2094	similar to the malic/Na^+-lactate antiporter MleN in *B. subtilis* (*40*)	amino acid change A94T caused by a nucleotide substitution G→A at the 280th nucleotide position in the gene
17	INT DIV	between divergently transcribed genes SA2125 and SA2126	SA2125 matches family consisting of arginases, agmatinases, and formiminoglutamases.	The underlined T in the potential sigma-A site **TTTATC**TCTCGGCTTGTAATA**TGTATAAT** was substituted with a C 237- and 53-bp upstream of SA2125 and SA2126 respectively. SA2126 is underexpressed by 13-fold in JH9 compared to JH1 (*41*).
18	SYN	SA2320 (*pfoR*)	contains a match to a domain of a sugar specific permease	nucleotide substitution T→C at the 504th nucleotide position in the 168th codon
(e) Appeared first in JH9				
(i) Loci previously implicated in both vancomycin and β-lactam resistance				
19	INT DIV	between divergently transcribed genes SA0526 and SA0527 (*nagB*)	NagB isomerizes glucosamine-6-P to fructose-6-P. Glucosamine-6-P occupies a central position between cell wall synthesis and glycolysis (*42*).	nucleotide substitution G→A 107- and 172-bp upstream of SA0526 and *nagB* respectively
(ii) Loci previously implicated in vancomycin resistance				
20	NONSYN	SA0617 (*vraG*)	codes for permease of an ABC transporter (*6*)	amino acid change A580V caused by a nucleotide substitution C→T at the 1739th nucleotide position in the gene
(iii) Other loci				
21	INT CONV	between convergently transcribed genes SAP007 and SAP008 (on plasmid)	SAP008 matches a family of alcohol dehydrogenases.	nucleotide deletion of an A 35- and 137-bp downstream of SAP007 and SAP008 respectively, occurred in a homopolymeric tract GAAAAAAT
22	FRAME	SA0171 (*fdh*)	matches family of D-isomer specific 2-hydroxyacid dehydrogenases	A nucleotide deletion of a T at the 17th nucleotide position in the gene frameshifted the last 98% of the gene. The deletion occurred in a homopolymeric tract GTTTTTTTTG.
23	NONSYN	SA0185	likely in an operon with two genes coding for components of the phosphoenolpyruvate::sugar phosphotransferase system	amino acid change A25D caused by a nucleotide substitution C→A at the 74th nucleotide position in the gene
24	NONSYN	SA0215	likely in an operon with a gene similar to periplasmic-iron binding protein BitC in *B. hyodynsenteriae* (*43*)	amino acid change D197G caused by a nucleotide substitution A→G at the 590th nucleotide position in the gene
25	SYN	SA0388 (*set12*)	codes for exotoxin	nucleotide substitution T→C at the 663rd nucleotide position in the 221st codon
26	INT TAND	between tandemly transcribed genes SA0557 and SA0558	SA0557 matches family consisting of a number of monomeric NADPH-dependent oxidoreductases.	nucleotide substitution T→C 321-bp upstream of SA0558 and 125-bp downstream of SA0557
27	NONSYN	SA1147	contains match to family of restriction endonucleases	amino acid change T9A caused by a nucleotide substitution A→G at the 25th nucleotide position in the gene
28	SYN	SA1510 (*gapB*)	codes for glyceraldehyde-3-phosphate dehydrogenase involved in gluconeogenesis	nucleotide substitution T→C at the 609th nucleotide position in the 203rd codon
29	NONSYN	SA1659 (*prsA*)	codes for a chaperone that assists post-translocational folding of proteins at the cytoplasmic/cell wall interface (*34-39*)	amino acid change E114Q caused by a nucleotide substitution G→C at the 340th nucleotide position in the gene
30	SYN	SA2091		nucleotide substitution A→G at the 693 nucleotide position in the 231st codon
31	SYN	SA2119	matches family of dehydrogenases	nucleotide substitution A→G at the 741st nucleotide position in the

				247th codon
32	INT TAND	between tandemly transcribed genes SA2232 and SA2233	SA2232 matches family of reductases involved in the alternative pyrimidine biosynthetic pathway. SA2233 has similarity to methylenomycin A resistance protein Mmr in *B. subtilis*.	nucleotide substitution T→C 144-bp upstream of SA2232 and 82-bp downstream of SA2233
33	TRNA	SAtRNA34	tRNA-Tyr	A nucleotide insertion of a C occurred at the 72nd nucleotide position of the tRNA with a length of 81 nucleotides. The insertion occurred in a homopolymeric tract GCCCCCCT.
(f) Unconfirmed§				
34	FRAME	in IS*1811* insertion sequence directly downstream of SA0617 (*vraG*)	VraG is a permease of an ABC transporter previously implicated in vancomycin resistance. See mutation 20.	putative deletion in JH9 of an A
35	INT CONV	between convergently transcribed genes SA2015 and SA2016 (*rpsI*)	RpsI is the 30S subunit of ribosomal protein S9.	putative deletion in JH9 of an A 26- and 150-bp downstream of SA2015 and SA2106 respectively, occurred in a homopolymeric tract CAAAAAAAAAAAAAG

In contrast to Table 2 in the main text, this table describes all 35 point mutations and gives the exact positions of the mutations in nucleotide coordinates and many more references to relevant research articles. In total, there were six frameshifts (mutations 1, 9, 10, 15, 22, and 34), one substitution that introduced a stop codon truncating a protein (mutation 11), 14 nonsynonymous substitutions (mutations 2-8, 14, 16, 20, 23, 24, 27, and 29), five mutations in intergenic sequence between divergently or tandemly transcribed genes (mutations 12, 17, 19, 26, and 32), one insertion in a tRNA (mutation 33), six synonymous substitutions (mutations 13, 18, 25, 28, 30, and 31), and finally two deletions between convergently transcribed genes (mutations 21 and 35). The 35 mutations fall into 31 separate loci, with the gene *rpoB* harboring four mutations (mutations 3-6) and the gene *prsA* harboring two mutations (mutations 15 and 29).

*, The numeric identifiers for the mutations used in Table 1 in the main text. An identifier printed in red indicates an insertion or deletion in a homopolymeric tract of initial length ≥ 6 nucleotides (e.g. aaaaaa).

†, The mutations are typed as follows: FRAME, a mutation causing a frameshift; INT CONV, a mutation in intergenic sequence between convergently transcribed genes; INT DIV, a mutation in intergenic sequence between divergently transcribed genes; INT TAND, a mutation in intergenic sequence between tandemly transcribed genes; NONSYN, a nonsynonymous substitution; SYN, a synonymous substitution; TRNA, a mutation in a reading frame coding for a tRNA; TRUNC, a mutation that truncated a gene.

‡, A locus is on the chromosome unless otherwise indicated. If a locus is on the plasmid, the box is shaded grey.

§, Mutations 34 and 35 could not be confirmed because of failure of the PCR sequencing method. Unique PCR primers could not be designed to amplify the region containing mutation 34 in the IS*1811* insertion sequence with multiple copies on the chromosome, and the sequencing reaction would not extend past the run of adenines containing mutation 35.

Table S.2. Descriptions of all 35 point mutations grouped by possible functions. See notes beneath table.

#*	Type of mutation†	Mutated locus‡	Description of mutated locus	Mutation(s)	N315	JH1	JH2	JH5	JH6	JH9	JH14	JH15
(a) Confirmed by PCR sequencing												
(I) Loci previously implicated in resistance to both vancomycin and β-lactams												
2	NONSYN	SA1702	in operon with gene *vraR*, which is possibly involved in regulation of cell wall synthesis (*4-7*)	amino acid change H164R caused by a nucleotide substitution A→G at the 491st nucleotide position in the gene								
19	INT DIV	between divergently transcribed genes SA0526 and SA0527 (*nagB*)	NagB isomerizes glucosamine-6-P to fructose-6-P. Glucosamine-6-P occupies a central position between cell wall synthesis and glycolysis (*42*).	nucleotide substitution G→A 107- and 172-bp upstream of SA0526 and *nagB* respectively								
(II) Loci previously implicated in resistance to vancomycin												
10	FRAME	SA1843 (*agrC*)	in the *agr* locus involved in qorum sensing and regulation of the expression of virulence and cell surface proteins (*17-25*)	A nucleotide deletion of a T at the 313th nucleotide position in the gene frameshifted the last 70% of the gene. The deletion occurred in a homopolymeric tract ATTTTTTTA.								
20	NONSYN	SA0617 (*vraG*)	codes for permease of an ABC transporter (*6*)	amino acid change A580V caused by a nucleotide substitution C→T at the 1739th nucleotide position in the gene								
(III) Loci involved in resistance to β-lactams												
1	FRAME	SAP011 (*blaR1*) (on plasmid)	involved in regulation of β-lactamase gene *blaZ* and broad spectrum β-lactam resistance gene *mecA* (*1-3*)	A nucleotide deletion of an A at the 466th nucleotide position in the gene frameshifted the last 70% of the gene. The deletion occurred in a homopolymeric tract TAAAAAAAAT.								
(IV) Loci involved in resistance to rifampin												
3	NONSYN	SA0500 (*rpoB*)	codes for β-subunit of RNA polymerase (*8-15*)	amino acid change D471Y caused by a nucleotide substitution G→T at the 1411th nucleotide position in the gene								
4	NONSYN			amino acid change A473S caused by a nucleotide substitution G→T at the 1417th nucleotide position in the gene								
5	NONSYN			amino acid change A477S caused by a nucleotide substitution G→T at the 1429th nucleotide position in the gene								
6	NONSYN			amino acid change E478D caused by a nucleotide substitution G→T at the 1434th nucleotide position in the gene								
(V) Loci previously implicated in resistance to daptomycin												
7	NONSYN	SA0501 (*rpoC*)	codes for β'-subunit of RNA polymerase (*11, 16*)	amino acid change E854K caused by a nucleotide substitution G→A at the 2560th nucleotide position in the gene, may map to outer rim of secondary channel (*16*)								
11	TRUNC	SA0019 (*yycH*)	in a gene cluster with the gene *yycF* and possibly involved in the regulation of the autolysin gene *lytM* (*11, 26-29*)	A nucleotide substitution G→A at the 107th nucleotide position in the gene converted the 36th codon to a stop codon, possibly truncating protein to 10% of its length.								
(VI) Other genes with changes that altered the protein												
8	NONSYN	SA1129	contains match to RNA binding motif	amino acid change D296Y caused by a nucleotide								

				substitution G→T at the 886th nucleotide position in the gene									
9	FRAME	SA1249	putative gene possibly in an operon with the gene *murG*, which is involved in cell wall synthesis	A nucleotide deletion of a G at the 31st nucleotide position in the gene frameshifted the last 80% of the gene. The deletion occurred in a homopolymeric tract TGGGGGGGGGA.									
14	NONSYN	SA0980 (*isdE*)	involved in passage of heme-iron to cytoplasm during pathogenesis (*31-33*)	amino acid change A84V caused by a nucleotide substitution C→T at the 251st nucleotide position in the gene									
15	FRAME	SA1659 (*prsA*)	codes for a chaperone that assists post-translocational folding of proteins at the cytoplasmic/cell wall interface (*34-39*)	A nucleotide deletion of an A at the 804th nucleotide position in the gene frameshifted the last 15% of the gene. The deletion occurred in a homopolymeric tract CAAAAAAAT.									
29	NONSYN			amino acid change E114Q caused by a nucleotide substitution G→C at the 340th nucleotide position in the gene									
16	NONSYN	SA2094	similar to the malic/Na^+-lactate antiporter MleN in *B. subtilis* (*40*)	amino acid change A94T caused by a nucleotide substitution G→A at the 280th nucleotide position in the gene									
22	FRAME	SA0171 (*fdh*)	matches family of D-isomer specific 2-hydroxyacid dehydrogenases	A nucleotide deletion of a T at the 17th nucleotide position in the gene frameshifted the last 98% of the gene. The deletion occurred in a homopolymeric tract GTTTTTTTTG.									
23	NONSYN	SA0185	likely in an operon with two genes coding for components of the phosphoenolpyruvate::sugar phosphotransferase system	amino acid change A25D caused by a nucleotide substitution C→A at the 74th nucleotide position in the gene									
24	NONSYN	SA0215	likely in an operon with a gene similar to periplasmic-iron binding protein BitC in *B. hyodynsenteriae* (*43*)	amino acid change D197G caused by a nucleotide substitution A→G at the 590th nucleotide position in the gene									
27	NONSYN	SA1147	contains match to family of restriction endonucleases	amino acid change T9A caused by a nucleotide substitution A→G at the 25th nucleotide position in the gene									
(VII) Regions coding for non-mRNAs													
33	TRNA	SAtRNA34	tRNA-Tyr	A nucleotide insertion of a C occurred at the 72nd nucleotide position of the tRNA with a length of 81 nucleotides. The insertion occurred in a homopolymeric tract GCCCCCCT.									
(VIII) Other intergenic regions between divergently and tandemly transcribed genes													
12	INT DIV	between divergently transcribed genes SAS014 and SA0411 (*ndhF*)	NdhF is the F-subunit of NADH dehydrogenase.	nucleotide deletion of a T 579- and 452-bp upstream of SAS014 and SA0411 respectively, occurred in a homopolymeric tract ATTTTTTTC									
17	INT DIV	between divergently transcribed genes SA2125 and SA2126	SA2125 matches family consisting of arginases, agmatinases, and formiminoglutamases.	The underlined T in the potential sigma-A site **TTTATC**TCTCGGCTTGTAATA**TGTATAAT** was substituted with a C 237- and 53-bp upstream of SA2125 and SA2126 respectively. SA2126 is underexpressed by 13-fold in JH9 compared to JH1 (*41*).									
26	INT TAND	between tandemly transcribed genes	SA0557 matches family consisting of a number of monomeric NADPH-dependent oxidoreductases.	nucleotide substitution T→C 321-bp upstream of SA0558 and 125-bp downstream of SA0557									

		SA0557 and SA0558											
32	INT TAND	between tandemly transcribed genes SA2232 and SA2233	SA2232 matches family of reductases involved in the alternative pyrimidine biosynthetic pathway. SA2233 has similarity to methylenomycin A resistance protein Mmr in *B. subtilis*.	nucleotide substitution T→C 144-bp upstream of SA2232 and 82-bp downstream of SA2233									
(IX) Genes with synonymous substitutions													
13	SYN	SA0582	similar to Na^+/H^+ antiporter subunit MrpE in *B. subtilis* (*30*)	nucleotide substitution T→C at the 90th nucleotide position in the 30th codon									
18	SYN	SA2320 (*pfoR*)	contains a match to a domain of a sugar specific permease	nucleotide substitution T→C at the 504th nucleotide position in the 168th codon									
25	SYN	SA0388 (*set12*)	codes for exotoxin	nucleotide substitution T→C at the 663rd nucleotide position in the 221st codon									
28	SYN	SA1510 (*gapB*)	codes for glyceraldehyde-3-phosphate dehydrogenase involved in gluconeogenesis	nucleotide substitution T→C at the 609th nucleotide position in the 203rd codon									
30	SYN	SA2091		nucleotide substitution A→G at the 693 nucleotide position in the 231st codon									
31	SYN	SA2119	matches family of dehydrogenases	nucleotide substitution A→G at the 741st nucleotide position in the 247th codon									
(X) Intergenic regions between convergently transcribed genes													
21	INT CONV	between convergently transcribed genes SAP007 and SAP008 (on plasmid)	SAP008 matches a family of alcohol dehydrogenases.	nucleotide deletion of an A 35- and 137-bp downstream of SAP007 and SAP008 respectively, occurred in a homopolymeric tract GAAAAAAT									
(b) Unconfirmed§													
34	FRAME	in IS*1811* insertion sequence directly downstream of SA0617 (*vraG*)	VraG is a permease of an ABC transporter previously implicated in vancomycin resistance. See mutation 20.	putative deletion in JH9 of an A	NOT DETERMINED								
35	INT CONV	between convergently transcribed genes SA2015 and SA2016 (*rpsI*)	RpsI is the 30S subunit of ribosomal protein S9.	putative deletion in JH9 of an A 26- and 150-bp downstream of SA2015 and SA2106 respectively, occurred in a homopolymeric tract CAAAAAAAAAAAAAAG	NOT DETERMINED								

In contrast to Table 2 in the main text, this table describes all 35 point mutations and gives the exact positions of the mutations in nucleotide coordinates and many more references to relevant research articles. The mutations have been grouped into categories according to possible functions (e.g. mutations in loci previously implicated in antibiotic resistance, nonsynonymous substitutions, synonymous substitutions, etc.). If a mutation appears in a given isolate, then the box for the isolate is shaded in grey. For an explanation of some of the conventions used in this table, see the notes below. In total, there were six frameshifts (mutations 1, 9, 10, 15, 22, and 34), one substitution that introduced a stop codon truncating a protein (mutation 11), 14 nonsynonymous substitutions (mutations 2-8, 14, 16, 20, 23, 24, 27, and 29), five mutations in intergenic sequence between divergently or tandemly transcribed genes (mutations 12, 17, 19, 26, and 32), one insertion in a tRNA (mutation 33), six synonymous substitutions (mutations 13, 18, 25, 28, 30, and 31), and finally two deletions between convergently transcribed genes (mutations 21 and 35). The 35 mutations fall into 31 separate loci, with the gene *rpoB* harboring four mutations (mutations 3-6) and the gene *prsA* harboring two mutations (mutations 15 and 29).

*, The numeric identifiers for the mutations used in Table 1 in the main text. An identifier printed in red indicates an insertion or deletion in a homopolymeric tract of initial length ≥ 6 nucleotides (e.g. aaaaaa).

†, The mutations are typed as follows: FRAME, a mutation causing a frameshift; INT CONV, a mutation in intergenic sequence between convergently transcribed genes; INT DIV, a mutation in intergenic sequence between divergently transcribed genes; INT TAND, a mutation in intergenic sequence between tandemly transcribed genes; NONSYN, a nonsynonymous substitution; SYN, a synonymous substitution; TRNA, a mutation in a reading frame coding for a tRNA; TRUNC, a mutation that truncated a gene.

‡, A locus is on the chromosome unless otherwise indicated. If a locus is on the plasmid, the box is shaded grey.

§, Mutations 34 and 35 could not be confirmed because of failure of the PCR sequencing method. Unique PCR primers could not be designed to amplify the region containing mutation 34 in the IS*1811* insertion sequence with multiple copies on the chromosome, and the sequencing reaction would not extend past the run of adenines containing mutation 35.

Table S.3. Differences found between the N315 and JH1 chromosomes.

(a) 3 replacements			
A 42,900-bp phage-like element φN315 in N315 is replaced in JH1 by an element with 70% identity. Two distinct 306- and 79-bp regions in N315 are replaced in JH1 by respectively 197- and 29-bp regions with no homology.			
(b) 82 local insertions/deletions not considering regions in part (a)			
41 insertions in JH1		**41 deletions in JH1**	
Size (bp)	**Frequency**	**Size (bp)**	**Frequency**
≥ 40,000	3*	≥ 40,000	0
10,000 to 39,999	0	10,000 to 39,999	1‡
1,000 to 9,999	4†	1,000 to 9,999	7‡
100 to 999	5	100 to 999	9
10 to 99	9	10 to 99	9
2 to 9	5	2 to 9	4
1	15	1	11
(c) 445 substitutions not considering regions in part (a)			
The 445 substitutions are scattered across the chromosome: only 12, 44, 64, and 154 are separated from the nearest substitution to the left or right by respectively ≤ 0-, 10-, 100-, and 1000-bp.			

*, All phage-like elements.

†, All involve mobile elements. Includes insertions of three >1000-bp IS*1811* insertion sequences.

‡, All involve mobile elements. Includes deletion of 6343-bp region of SCC*mec* cassette that includes bleomycin resistance gene *bleO* but not *mecA*, *mecR1*, or *mecI*. Also includes deletions of 15,659-bp pathogenicity island SaPIn1, three 6712-bp Tn*554* transposons, and three >1,000-bp IS*1811* insertion sequences.

Table S.4. The *vraR* operon was PCR sequenced in seven pairs of isolates, each consisting of a non-VISA and a closely related VISA.

Strain	Source	Vancomycin MIC (μg/ml)	MLST*	*spaA* type†	PFGE pattern‡	Point mutation rate for a region like that PCR sequenced§	Reference(s)
Non-VISA/VISA pair 1							
JH1	clinical (New York)	1.0	1-4-1-4-12-1-28	TJMBMDMGMK	A0	<1:80,000¶	(*39*, *44*)
JH9		8.0	1-4-1-4-12-1-28	TJMBMDMGMK	A0		
Pair 2							
N315	clinical (Japan)	0.5	1-4-1-4-12-1-10	TJMBMDMGMK	A1	<1:5,000¶	(*45*)
MU50		8.0	1-4-1-4-12-1-10	TJMBMDMGMK	A2		
Pair 3							
PC1	clinical (New York)	2.0	1-4-1-4-12-1-10	TJMBMDMGMK	A3	<1:80,000‖	(*46*)
PC3		8.0	1-4-1-4-12-1-10	TJMBMDMGMK	A3		
Pair 4							
N315	clinical (Japan)	0.5	1-4-1-4-12-1-10	TJMBMDMGMK	A1	<1:3,000**	(*45*, *47*)
VNJ	clinical (New Jersey)	8.0	1-4-1-4-12-1-10	TMBMDMGMK	A4		
Pair 5							
N315	clinical (Japan)	0.5	1-4-1-4-12-1-10	TJMBMDMGMK	A1	<1:3,000**	(*45*, *47*)
VMI	clinical (Michigan)	8.0	1-4-1-4-12-1-10	TJMGMK	A5		
Pair 6							
E-MRSA15	clinical (England)	0.75	7-6-1-5-8-8-6	TJJEJNF2MNF2MOMOKR	A6	<1:3,000**	unpublished (H.M.L.)
HSMB1	clinical (Portugal)	4.0	7-6-1-5-8-8-6	TJJEJNF2MNF2MOMOKR	A7		
Pair 7							
COL	clinical (England)	1.5	3-3-1-1-4-4-16	MK	A8	≤1:1000,000††	(*48*)
VM3	laboratory (Tomasz)	6.0	3-3-1-1-4-4-16	MK	A8		

On the N315 chromosome, the precise region examined was the 3,090-bp region that has the coordinates 1,946,630-1,949,719 and includes the four genes *vraR*–*vraS*–SA1702–SA1703 that make up the *vraR* operon (*4*). Each of the seven pairs of isolates was made up of a non-VISA (vancomycin MIC < 4 μg/ml) and a closely related VISA (MIC ≥ 4 μg/ml).

***,** The multi-locus sequence type (MLST) of an isolate is based on the sequences of seven housekeeping genes (*49*).

†, The *spaA* type is based on the polymorphic region of protein A consisting of 24-bp repeats. The diversity of this region arises from the duplication and deletion of the repeats (*50*).

‡, The pulse field gel electrophoresis (PFGE) pattern was obtained by digesting chromosomal DNA with *Sma*I endonuclease and separating the DNA fragments by PFGE.

§, The rate of point mutations between the non-VISA and VISA in a region like the 3,090-bp region that was PCR sequenced. The 3,090-bp region that was PCR sequenced consisted of the *vraR* operon and a little flanking intergenic sequence. The rate of point mutations in such a region would be expected to be less than say the rate for a completely intergenic region.

¶, The rate of point mutations chromosome-wide, computed by comparing the whole chromosomal sequences (see (*45*) for the N315 and Mu50 sequences). Since this rate includes point mutations in intergenic sequence and pseudogenes, it is considered to be an overestimate of the rate in a primarily coding region like the 3,090-bp region that was PCR sequenced.

‖ For PC1 and PC3, the rate is expected to be similar to that between JH1 and JH9. Like JH1 and JH9, PC1 and PC3 differ by about three months of *in vivo* evolution and belong to a series of isolates taken from a patient being treated with vancomycin that have identical MLSTs, *spaA* types, and PFGE patterns (*46*).

******, A rate was estimated from the MLSTs. The two isolates have identical MLSTs and so are identical in the 3,198-bp of sequence tested in the seven housekeeping genes used for MLST (*49*). Thus, the rate is <1:3,000. For our purposes, it was valid to estimate a rate from the MLSTs because the regions used for MLST were found to be no less polymorphic than the 3,090-bp region that was PCR sequenced. When we examined the multialignments of orthologous regions from sequenced non-VISA strains (N315, MW2, 476, USA300, COL, 8325, and E-MRSA252) (*51*), about 1.2% of the columns in the multi-alignment contained differences between the strains in the case of the regions used for MLST compared to only 1.0% of the columns containing differences in the case of the 3,090-bp region that was PCR sequenced.

††, COL and VM3 are expected to differ by only a few mutations. The vancomycin susceptible isolate COL was grown overnight in an antibiotic free medium and plated on an agar plate containing 3.0 μg/ml of vancomycin, and a mutant colony capable of growing on the plate in the presence of vancomycin was picked in a single step and called VM3 (*48*). Consequently, COL and VM3 are expected to differ by perhaps only one mutation and by at most a few mutations on the chromosome, which in COL has a length of roughly 2.9-Mbp (*52*). To see why this should be, consider one estimate that bacteria have a spontaneous mutation rate of about 1/300 per genome per replication (includes single nucleotide substitutions, insertions, and deletions as well as larger mutations) (*53*). For COL, this translates to a rate of about 10^{-9} spontaneous mutations per base pair per division. COL and VM3 are separated by about 12 hours of *in vitro* evolution or about 24 doublings, so one might expect the number of mutations in VM3 to be roughly $10^{-9} \times 2.9 \times 10^{6} \times 24 = 0.07$.

Supporting Figures

Figure S.1. Differences found between N315 and JH1. (a) On chromosome. Shown is a schematic of the 2,967,572 column alignment of the N315 and JH1 chromosomes. The positions of detected differences are marked: **red**, indicates an insertion in JH1 relative to N315; **blue**, indicates a deletion in JH1 relative to N315; and **green**, indicates a nucleotide substitution. The insertions and deletions range in size from 1-bp (narrowest of bands) to >40,000-bp (widest of bands). Three replacements involving distantly related or non-homologous sequence produced dense clusters of polymorphisms (**aqua**). The largest replacement involved the swapping of a ≈40,000-bp phage-like element in N315 with an element in JH1 with only 70% base pair identity. Excluding the three regions of replacement, we detected 82 insertions and deletions and 445 nucleotide substitutions chromosome-wide. Included in this list are insertions in JH1 of three >40,000-bp phage-like elements and a deletion in JH1 of a 15,659-bp pathogenicity island (**orange**). **(b) On plasmid.** The 30,429-bp JH1 plasmid is composed of segments bearing >99% identity to either the 24,653-bp N315 plasmid or the 25,107-bp Mu50 plasmid. The single segment with similarity to the N315 plasmid spans almost half the plasmid (**black**). Four segments are similar to the transposable element *tnpE* on the Mu50 plasmid (**brown**). Additionally, three other segments are similar to the Mu50 plasmid (**gray**). The remaining segments have homology to neither the N315 nor Mu50 plasmids and code for genes of unknown function (white).

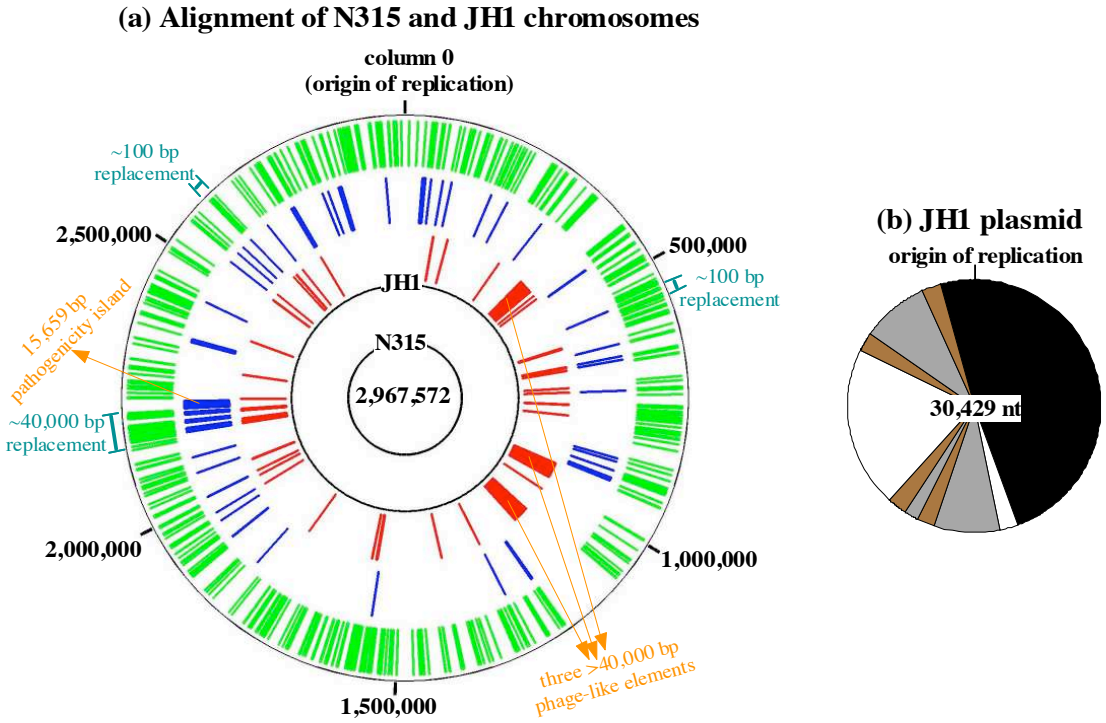
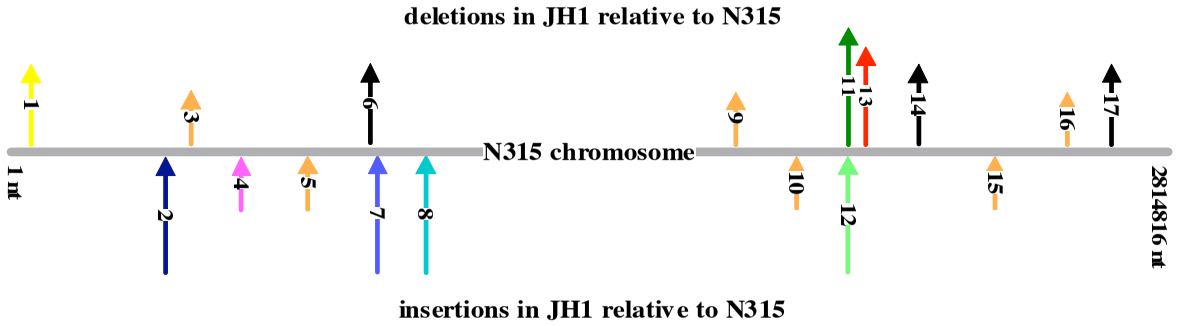


Figure S.2. Differences ≥ 1000-bp found between the N315 and JH1 chromosomes. The N315 chromosome is shown. Arrows above the chromosome indicate deletions in JH1 relative to N315. Arrows below indicate insertions in JH1 relative to N315. The length of an arrow is proportional to the logarithm of the size of the insertion or deletion. The inserted and deleted elements are color coded and numbered. The first table describes the types of elements (center of page). The second table gives the positions of the elements (bottom of page).



Color code for chromosomal elements numbered 1-17 above.

Color	Type of element (reference and GenBank accession number given)	Size (bp)	Number of copies (≥ 95% similarity) in N315	strains with homologs (≥ 50% similarity)*							Examples of resistance, toxin, and virulence genes on element
				252	476	8325	COL	MW2	N315	Mu50	
	Region of type II SCCMec cassette (*54*) (D86934), includes bleomycin resistance gene *bleO* **but not** *mecA*, *mecR1*, or *mecI*	6,343	1	■					■	■	bleomycin resistance gene *bleO*
	pathogenicity island SaPIn1 (*45*) (BA000018)	15,659	1						■	■	toxin shock syndrome toxin-1 gene *tst*
	> 50% similarity to phage 92 (*55*) (AY954967)	>40,000	0								
	> 50% similarity to phage φ11 of 8325 (*56*) (AF424781)	44,089	0			■				■	
	> 60% similarity to phage ΦSA2usa of USA300 (*57*) (CP000255)	45,503	0	■	■	■	■	■			virulence associated protein E gene *virE*
	phage φN315 (*45*) (BA000018)	43,800	1	■					■	■	enterotoxin P gene *sep*
	> 70% similarity to phage φN315 (*45*) (BA000018)	>40,000	0	■					■	■	
	Tn554 transposon (*58*) (X03216)	6,712	5	■					■	■	erythromycin and spectinomycin resistance genes *ermA* and *spc*
	IS1181 insertion sequence (*59, 60*) (L14544)	1,520	8			■	■		■	■	
	cluster of rRNAs and tRNAs	>1,000	several	■	■	■	■	■	■	■	

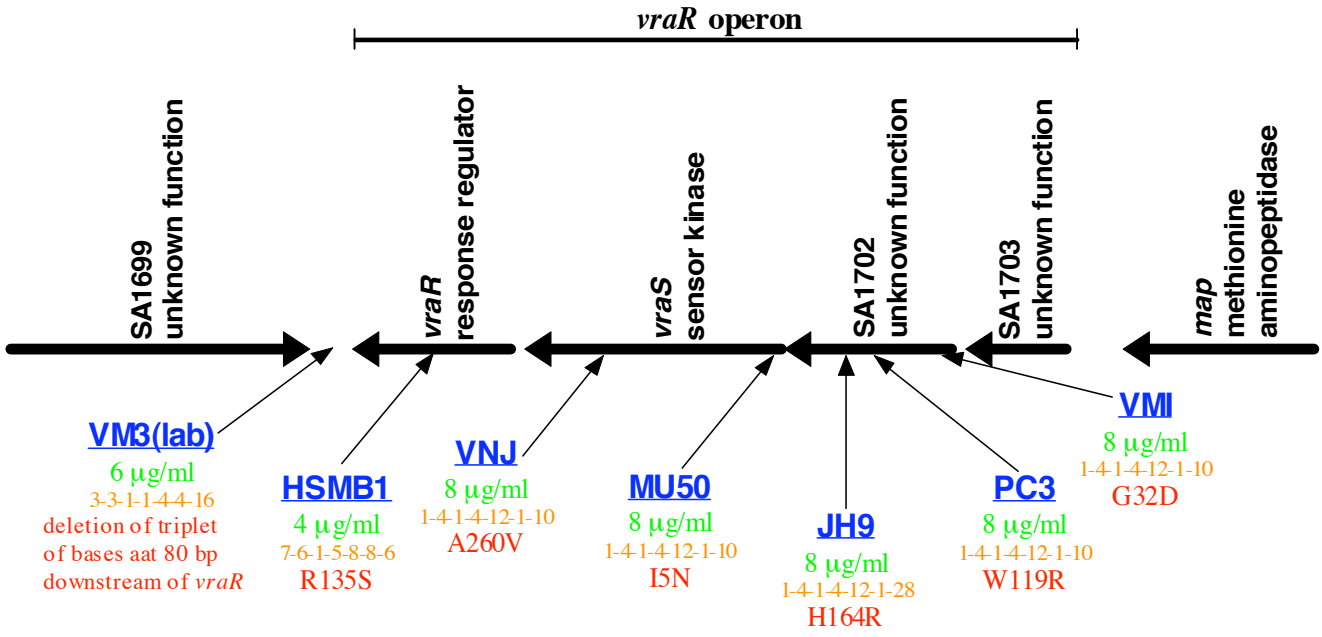
*Only Mu50 is a VISA. The remainder are VSSAs.

Positions of chromosomal elements numbered 1-17 above.

Number	Type[a]	N315 position begin	N315 position end	Number	Type[a]	N315 position begin	N315 position end	Number	Type[a]	N315 position begin	N315 position end	Number	Type	N315 position begin	N315 position end
1	D	37166	43508	5	I	712808	712809	9	D	1761590	1763109	14	D	2198763	2205474
2	I	367374	367375	6	D	866875	873586	10	I	1901485	1901486	15	I	2378386	2378387
3	D	426553	428072	7	I	885643	885644	11/12[b]	R	2005721	2049520	16	D	2566722	2568241
4	I	553019	553020	8	I	997087	997088	13	D	2056679	2072337	17	D	2671105	2677816

[a]I = insertion in JH1 relative to N315, D = deletion in JH1 relative to N315, and R = replacement
[b]Element 11 was replaced by 12.

Figure S.3. Mutations in the *vraR* operon in all seven VISA isolates. As discussed in Table S.4, a 3,090-bp segment encompassing the *vraR* operon *vraR*–*vraS*–SA1702–SA1703 was PCR sequenced in seven pairs of *S. aureus* isolates, with each pair consisting of a non-VISA and a closely related VISA [names, vancomycin MICs, and MLSTs (*49*) of the VISAs in blue, green, and orange respectively]. For each pair of isolates, we: (i) found a difference in the 3,090-bp segment between the non-VISA and VISA, (ii) confirmed that this difference is real by examining both the forward and reverse PCR sequences and traces, and (iii) showed that the difference is due to a mutation (red) in the VISA by considering outlying more distantly related sequenced *S. aureus* isolates (*51*). In each of the six VISAs JH9, MU50, PC3, VNJ, VMI, and HSMB1, a single nonsynonymous substitution was observed that resulted in the indicated amino acid change (red) in either VraR, VraS, or SA1702. In the VISA VM3, the only mutation observed was a deletion (red) of a triplet of bases aat 80-bp downstream of *vraR*. Since mutations are always to be expected in a sufficiently large locus between sufficiently divergent isolates, it was necessary to assess the statistical significance of the observation that the 3,090-bp segment encompassing the *vraR* operon is mutated in all the VISAs. Of the differences seen, mutations in three of the VISAs are particularly significant. As indicated in Table S.4, the VISAs JH9, PC3, and VM3 are very closely related to their parental non-VISAs JH1, PC1, and COL respectively (see in particular the point mutation rates in column 7 in Table S.4 and the notes at the bottom of the table). That mutations were found in the 3,090-bp segment in all three VISAs and not their parental non-VISAs is likely not a chance event. Though the mutation in VM3 is between convergently transcribed genes, it too is believed to be significant. VM3 was obtained by growing the vancomycin susceptible isolate COL overnight in an antibiotic free medium, plating the culture on an agar plate containing 3 μg/ml of vancomycin, and picking in a single step a mutant colony that could grow on the plate in the presence of the antibiotic (*48*). Hence, VM3 is expected to differ from its parent COL by perhaps only one mutation and by at most a few mutations on the chromosome, which has a length of roughly 2.9-Mbp (see footnote †† at bottom of Table S.4). That one of these mutations would happen to fall very near the *vraR* operon by chance is extremely improbable. The observation that the 3,090-bp segment is mutated in all seven of the VISAs examined was shown to be highly statistically significant. To show statistical significance, we computed the probability *P* that one would find at least one 3090 segment anywhere on the chromosome to be point mutated in all the VISAs by chance. The null model used took into account the point mutation rates between the isolates (column 7 in Table S.4) and hypothesized that the point mutations occur chromosome-wide in the isolates according to a uniform spatial distribution. The probability *P* was found to be less (possibly much less) than 0.001, which means no 3,090-bp region anywhere on the chromosome should be expected to be mutated in all the VISAs by chance. Two possibilities could account for the fact that the 3,090-bp region encompassing the *vraR* operon was found to be mutated in all the VISAs: (i) the mutations observed in the locus underwent positive selection or (ii) the stretch of DNA is intrinsically polymorphic even in the absence of selection (e.g. insertions and deletions can occur in a homopolymeric tract with high frequency but not necessarily exert a phenotypic and therefore selectable effect). To rule out the latter possibility, we examined 10 other regions with similar fractions of coding sequence. In the seven sequenced non-VISA strains N315, MW2, 476, USA300, COL, 8325, and E-MRSA252 (51), the 3,090-bp region encompassing the *vraR* operon was not found to be especially polymorphic compared to the 10 other regions. Indeed, in the multi-alignments of the orthologous regions from the seven sequenced non-VISA strains, only 1.0% of the columns contained differences in the case of the 3,090-bp region encompassing the *vraR* operon compared to about 1.2% of the columns in the case of the concatenation of the regions in the seven housekeeping genes used for MLST (49).



Supporting Methods

1. Organization

We present our methods in three different levels of detail.

First, you may want to read Section 2 called "Sketch", in which we present a 2-3 page sketch of our methods. We expect the level of detail provided in this section will satisfy most readers.

If you would like more details than provided in Section 2, you may then want to read Section 3 called "More detailed summary". In this section, we provide a 4-5 page discussion of our methods.

If you while reading Section 3 are interested in learning even more details about a particular topic, you need not read the entire treatise. From Section 4 onwards, we elaborate on various topics, sometimes explaining terminology that may not be familiar to everyone. If you want more information about one of the topics, just skip to the appropriately named section (e.g. a detailed description of our Bayesian probabilistic model is provided in Section 11).

In general, this document was written to be comprehensive but also scannable so that desired information can be located quickly. Titles of sections and subsections are printed in bold font.

The numbers in parentheses [e.g. (*1*), (*4*, *5*), (*6*), etc.] do not refer to sections but rather to the references listed at the end of the text.

2. Sketch

Assessment of overrepresentation of mutations in homopolymeric tracts. The reported P-score (10^{-7}) was computed under a null model that point mutations occur according to a uniform spatial distribution irrespective of homopolymeric tracts.

Assessment of overlaps of lists of transcriptional changes. The reported P-score (10^{-11}, 10^{-2}, or 10^{-2}) is the probability of the degree of overlap of the list of genes controlled by the regulator and the list of genes differentially expressed in JH9 compared to JH1. The P-score was computed under a null model that the two lists are chosen independently. An effort was made to account for correlations in the expressions of genes in operons by considering predicted operons.

Assessment of the observation that mutations in the *vraR* operon were found in all seven of the VISAs examined. As detailed in Table S.4 and Figure S.3, we PCR sequenced the *vraR* operon in seven pairs of isolates, with each pair consisting of a non-VISA and a closely related VISA. For each pair, we: (i) found a difference in the *vraR* operon between the non-VISA and VISA, (ii) confirmed that this difference is real by examining both the forward and reverse PCR sequences and traces, (iii) and showed that this difference is due to a mutation in the VISA using outlying more distantly related sequenced *S. aureus* isolates (51). We wanted to ensure that the *vraR* operon is not just intrinsically polymorphic even in the absence of selective pressure. In seven sequenced non-VISAs (51), we compared the variation seen in the *vraR* operon with that seen in 10 other loci with similar fractions of coding sequence. The *vraR* operon was not found

to be especially polymorphic – in particular, it was found to be less polymorphic than the regions in the housekeeping genes used for MLST (49). To assess the significance of the observation that the *vraR* operon is mutated in all the VISAs, the appropriate P-score to compute is the probability of observing at least one segment the size of the *vraR* operon anywhere on the chromosome to be point mutated in all the VISAs by chance. The P-score can be computed by repeatedly starting with seven chromosomal sequences (the non-VISAs) and mutating the sequences (to get the artificial VISAs) by introducing point mutations according to a uniform spatial distribution, using the point mutation rates between the non-VISAs and VISAs (column 7 of Table S.4). The P-score can then be estimated as the frequency of the observation that somewhere on the chromosome there is at least one segment the size of the *vraR* operon that is point mutated in all the artificial VISAs. The P-score estimated in this way was found to be 0.001. The true P-score is expected to be less (possibly much less) than 0.001 for at least two reasons: (i) The upperbounds of the point mutation rates in Table S.4 were used. (ii) In cases where the non-VISA is not the parent of the VISA (N315 versus MU50, N315 versus VNJ, N315 versus VMI, and E-MRSA15 versus HSMB1), the point mutation rates that were used reflect mutations in both the non-VISA and VISA, which is tantamount to ignoring the added significance that the mutations were found to be exclusively in the VISAs.

High fidelity of N315 and MU50 sequences. After the two related isolates N315 and MU50 were sequenced in 2001 (included finishing) (45), many errors in the sequences were eliminated in 2004 when the sequences were compared and the hundreds of differences found were checked by PCR sequencing (61).

Sequencing and *de novo* assembly. For each of JH1 and JH9, the whole genome shotgun sequencing (62, 63) was carried out to a mean depth of 8.5-9.5X coverage, and the Celera assembler (64) was used to assemble *de novo* the paired end reads from the 3, 6, and 40 kb clone libraries, producing 62 JH1 and 79 JH9 contigs. For each of JH1 and JH9, one contig was the complete sequence of a circular 30 kb plasmid and the other contigs were partial sequences of a circular chromosome.

Comparison of plasmid sequences. The JH1 and JH9 contigs representing the complete plasmid sequences were compared and were found to differ over their entire lengths by only 3 isolated nucleotide differences. PCR sequencing confirmed two were real and showed the other was a read error.

Ordering of contigs and estimation of the sizes of the contig gaps. The JH1 and JH9 contigs representing chromosomal sequence were ordered using the N315 chromosomal sequence and the paired end reads bridging the contigs. With the exception of a few contig gaps in inserts specific to the JH lineage, the sizes of the contig gaps could be estimated using N315. Moreover, the sizes of 70-80% of the contig gaps could be estimated from paired end reads bridging contigs. A size estimate of a contig gap from paired end reads is a normally distributed random variable with a mean and standard deviation. In total, the size of every contig gap could be estimated from N315 and/or paired end reads. The JH1 and JH9 chromosomes are each estimated to be about 2.9 Mb long, a figure which includes the contig gaps. The JH1 and JH9 contig gaps are estimated to contain a mere 1.5 and 2.3% respectively of the JH1 and JH9 chromosomes.

Multi-alignment of the assembled N315, JH1, and JH9 chromosomal sequences and JH1 and JH9 reads. The whole genome multi-alignment program MGA (65) was used to produce a chromosome-wide multi-alignment (MAC) of the N315 chromosomal sequence and the ordered JH1 and JH9 contig chromosomal sequences. Thus, in each column of the MAC, there was a single base from each of N315, JH1, and JH9. The multi-alignment program ClustalW (66) was used in windows to augment the MAC with the mapped JH1 and JH9 reads and therefore produce a full chromosome-wide multi-alignment (MACR) of the assembled chromosomal sequences and reads.

Identification of large mutations. Large mutations between the N315, JH1, and JH9 chromosomes were searched for by very carefully examining the entire MAC. No large difference between JH1 and JH9 was found in the estimated 98.5 and 97.7% respectively of the JH1 and JH9 chromosomal sequence internal to the contigs. Moreover, it is unlikely that large differences between JH1 and JH9 exist in the contig gaps estimated to contain only 1.5 and 2.3% respectively of the JH1 and JH9 chromosomal sequence. With one exception, the means of the size estimates of the gaps from the pair end reads in one JH isolate always agreed within three standard deviations with the other JH isolate. In the single exceptional case, the gap was PCR sequenced to show that JH1 and JH9 were identical. When several other gaps thought most likely to harbor a large difference were PCR sequenced, JH1 and JH9 were again found to be identical.

Identification of small mutations. To find small mutations between the N315, JH1, and JH9 chromosomes, a Bayesian probabilistic model (BPM) was formulated to call nucleotide differences (NDs) in columns in the MACR. The BPM assumed that the N315 sequence contained no errors and considered the coverage and Phred quality values (67) of the JH1 and JH9 reads. Replacements, insertions, and deletions involving ≥ 2 bases were detected as clusters of NDs. If a ND in a column in the MACR was either predicted or ruled out by our BPM with a probability $P > T = 1\text{-}1/3000{,}000$, then the column was said to be informative. The stringent threshold T was selected to ensure that a ND was correctly predicted or ruled out in every informative column in the MACR without a single expected error. In other words, NDs would have been expected to have been identified with 100% accuracy if the analysis had been restricted to only informative columns. Such confidence could not be achieved when considering uninformative columns, which were almost exclusively in regions of 0 or 1X coverage and poor quality 2X coverage. In the N315 and JH1 comparison, 97% of the columns in the MACR were informative, and only NDs predicted in the informative columns were reported. Thus, no false ND between N315 and JH1 due to a read error in JH1 is expected to have been counted. In the JH1 and JH9 comparison, 94% of the columns in the MACR were informative. Every predicted ND in the informative columns was confirmed by PCR sequencing, except the deletion in the stretch of 14 adenines that could not be sequenced (Table 1 in the main text and Tables S.1 and S.2). Also considered were 10 of the more promising predictions of NDs in the uninformative columns. All but two were shown to be false by PCR sequencing. One proved to be real, and the other was the deletion in the IS*1811* insertion sequence that could not be PCR'd (Tables 1, S.1, and S.2). The NDs in the informative columns occurred at a rate of only 35 in 94% of the columns. Assuming that the NDs occurred at the same rate in uninformative columns, the number of NDs expected to have gone unreported in the 6% of uninformative columns in the MACR can be estimated as $35 / 0.94 \times 0.06 = 2.2$.. When a second more rigorous estimate was

done that considered the coverage and read quality in the uninformative columns, a number of 1.5 was obtained.

3. More detailed summary

Definition of ND. Henceforth, the term "ND" (short for nucleotide difference) refers to any difference between two sequences in a single column in the alignment of the two sequences. The term ND will refer not only to bona fide mutations but sequencing and assembly errors. A ND can involve an insertion, deletion, or substitution. When isolated, it can be a point mutation. However, it can also be part of a run of ND's arising due to a larger mutation. For example, a run of 1000 ND's would be produced by a single insertion of a 1000 nucleotides long element. We use the term "real ND" to refer to a ND that arose due to a bona fide mutation as opposed to a sequencing or assembly error.

Assessment of overrepresentation of mutations in homopolymeric tracts. The reported P-score (10^{-7}) is the probability of eight or more of the 33 confirmed point mutations falling into homopolymeric tracts of initial length ≥ 6 bp under a null model that point mutations occur according to a uniform spatial distribution irrespective of homopolymeric tracts. Homopolyermic tracts of length ≥ 6 bp comprise slightly less than 1.5% of the total sequence. The P-score is given by the bionomial distribution.

Assessment of overlaps of lists of transcriptional changes. A previous study identified the open reading frames (ORF's) (i.e. predicted genes) differentially expressed by ≥ 2-fold or more in JH9 compared to JH1 (41). We in this study observed mutations in JH9 in loci coding for the transcriptional regulators VraR, Agr, and YycF. Prior work had identified the ORF's controlled directly or indirectly by these three regulators (5, 18, 26), as well as the positive regulators TRAP (19) and ArlR (20) of the *agr* locus. The reported P-score (10^{-11}, 10^{-2}, or 10^{-2}) is the probability of overlap of the list of ORF's controlled by the regulator and the list of ORF's differentially expressed in JH9 compared to JH1. The P-score was computed under a null model that the two lists are chosen independently. The P-score is given by the Poisson distribution. An effort was made to account for correlations in the expressions of genes in operons by considering predicted operons.

Assessment of the observation that mutations in the *vraR* operon were found in all seven of the VISAs examined. The P-score's upperbound of 0.001 was calculated by computer simulation as follows: Let M_i denote the point mutation rate between the non-VISA and VISA in the i^{th} pair of isolates (see column 7 in Table S.4). To be conservative in our assessment of significance, we set the M_i to their upperbounds: M_i = 1:80,000, 1:5000, 1:80,000, 1:3000, 1:3000, 1:3000, or 1:1000,000 for i = 1 to 7 respectively (Table S.4). We started with the set S of all integers from 1 to 2,900,000. In a single trial, we for each i = 1 to 7 generated a subset S_i of integers by randomly selecting ceil($M_i \times 2900000$) different integers from S. Subject to the constraint that each S_i had to have ceil($M_i \times 2900000$) *unique* integers, integers were selected from S with equal likelihood. Here, selection does not mean removal, so S always remained unchanged. We determined if there existed an integer j satisfying $1 \leq j \leq 2900000 - 3090 + 1$

such that all subsets S_i for $i = 1$ to 7 contained an integer in the interval $[j, j+3090-1]$. The trial was deemed a success if such a j existed and a failure otherwise. After conducting 10,000 such trials, it was clear that successful trials occurred at a frequency of 0.001.

Sequencing and assembly statistics. For JH1, there are 62 contigs (39 with length ≥ 10,000 bp). For JH9, there are 79 contigs (48 with length ≥ 10,000 bp). In each of JH1 and JH9, one contig represents the complete sequence of a 30 kbp circular plasmid, and the remaining contigs represent parts of the sequence of a 2.9 Mbp circular chromosome. In JH1 and JH9, the chromosomal contigs can be grouped into about 15 and 23 scaffolds respectively (a scaffold is defined such that within a scaffold the order of contigs can be determined and the distances between contigs can be estimated using paired end reads bridging contigs without reference to another sequenced *S. aureus* strain). Order of all contigs could be determined and distances between all contigs could be estimated using paired end reads bridging contigs and/or N315. For JH1, 98.5% of the chromosome has a coverage ≥ 1X, with the coverage ranging from 0 to 26X and having a mean of 8.5X. For JH9, 97.7% of the chromosome has a coverage ≥ 1X, with the coverage ranging from 0 to 27X and having a mean of 9.5X.

Available for download or upon request. (a) Raw and trimmed JH1 and JH9 reads with base specific Phred quality values. (b) Complete JH1 and JH9 plasmid sequences. (c) Full multi-alignment of JH1, JH9, and N315 chromosomal sequences and trimmed JH1 and JH9 reads with site specific Phred quality values. (d) Various programs.

Multi-alignment programs used. clustalw (66). dialign (68). MGA, which is capable of globally multi-aligning closely related whole bacterial chromosomes (65). The .align file outputted by MGA is particularly useful since it is a succinct summary of all the differences between the sequences in the MGA global multi-alignment.

High fidelity of N315 and MU50 sequences. After the two related isolates N315 and MU50 were sequenced in 2001 (included finishing) (45), many errors in the sequences were eliminated in 2004 when the sequences were compared and the hundreds of differences found were checked by PCR sequencing (61).

Sequencing and *de novo* assembly of the JH1 and JH9 genomes and mapping of the JH1 and JH9 reads onto respectively the JH1 and JH9 contigs. Unless explicitly stated otherwise, the following applies to each of JH1 and JH9: The whole genome shotgun sequencing (62, 63) was done by the Joint Genomes Institute (69). Each base call in each read was assigned its own Phred quality value (67). The reads were trimmed for both vector and quality using a specialized trimming pipeline. The mean coverage of the trimmed reads was estimated to be 8.5-9.5X. The *de novo* assembly of the trimmed reads was done using the Celera assembler (64). Independently of the assembler, a mapping of the reads onto the contigs was generated using the q-gram technique (70). As is standard, some minimal editing was done to remove contigs arising from containment sequence. During this editing, <2% of the contig sequence was eliminated. The contigs that were discarded included contigs with >99% nucleotide identity to the sequenced *E. coli K12* and human genomes. For JH1 and JH9, there remained 62 and 79 contigs respectively.

Determination of the JH1 and JH9 plasmid and chromosomal sequence and ordering of the JH1 and JH9 contigs using N315. Previously, it was shown by MLST typing (49) that the JH isolates JH1-JH15 (each with a MLST 1-4-1-4-12-1-8) are closely related to the sequenced isolates N315 and Mu50 (both with a MLST 1-4-1-4-12-1-10) (44). When we used MGA to multi-align several randomly chosen large JH1 contigs with the complete N315 and Mu50 genomic sequences, it became clear that JH1 is more related to N315 than Mu50. From the multi-alignments, the point mutation rate between JH1 and N315 was crudely estimated to be 1:5000 bp. Unless explicitly stated otherwise, the following applies to each of JH1 and JH9: All the JH contigs were blasted against the N315 plasmid and chromosomal sequence. One JH contig was found to be a complete plasmid sequence. This contig exhibited high homology to the N315 plasmid and was circular. The remaining JH contigs were found to contain chromosomal sequence. The JH contigs could be ordered by position using read pairs bridging the contigs and the high homology to the N315 chromosome. No change in synteny between the JH and N315 chromosomes was observed, apart from several transpositions involving elements <10,000 bp. For each of JH1 and JH9, we therefore identified a complete plasmid sequence and produced an ordered set of contigs containing chromosomal sequence.

Identification and subsequent experimental verification of the mutations between the JH1 and JH9 plasmids. The program clustalw was used to align the complete JH1 and JH9 plasmid sequences. The two sequences were found to differ over their entire lengths by only 3 isolated ND's. PCR sequencing confirmed that two of the ND's were bona fide point mutations and showed that the other was a sequencing error.

Preliminary construction of the MAC. For each of JH1 and JH9, we concatenated in order the contigs containing chromosomal sequence, making sure to always place between two consecutive contigs a X to mark the contig gap. For each of JH1 and JH9, we therefore generated a single long chromosomal sequence punctuated with X's. We used MGA to construct a multi-alignment (referred to herein as the MAC) of these JH1 and JH9 chromosomal sequences and the N315 chromosomal sequence. Hence, each column in the MAC contained a single base from each of N315, JH1, and JH9.

Estimation of the sizes of the JH1 and JH9 contig gaps, the ruling out of large differences between JH1 and JH9 in the contig gaps, and the editing of the MAC. The .align file outputted by MGA was used to very carefully examine by eye the entire MAC. A JH1 or JH9 contig gap could be identified as an indel adjacent to a X. With the exception of several JH1 and JH9 contig gaps in new sequence specific to the JH lineage, the size of each gap could be inferred from the length of the corresponding region in N315. For 70-80% of the JH1 and JH9 contig gaps, the size of the gap could also be estimated from read pairs spanning the gap. Ultimately, the size of every JH1 and JH9 contig gap could be estimated from N315 and/or read pairs. While a size estimate from N315 is a single number, a size estimate from read pairs is a normally distributed random variable, with a mean and a standard deviation (STD). When all the estimates from N315 and all the means of the estimates from the read pairs were examined in both JH1 and JH9, it was found that >80% of the values were <1000 nucleotides and all were <10,000 nucleotides. When all the STD's of the estimates from the read pairs were examined in both JH1 and JH9, it was found that >80% of the values were <1000 nucleotides and all were <4000 nucleotides. In all but one case, the mean of the size estimate of a contig gap in one JH

strain computed from read pairs agreed to within three STD's with the lengths of the corresponding regions in N315 and the other JH strain. In most cases, the agreement was in fact to within one or two STD's. In the single exceptional case in which the disagreement exceeded three STD's, the sequence in the gap in JH1 was PCR sequenced to show that it was identical to the corresponding sequence in N315 and JH9. We also checked (e.g. by PCR sequencing) other cases of gaps in JH1 or JH9 thought most likely to harbor large differences, and in each case, JH1 and JH9 were found to be identical. Using the estimates of the sizes of the JH1 and JH9 contig gaps, the MAC was edited. Each X marking a JH1 or JH9 contig gap was replaced by a string of N's with a length equal to the estimate of the size of the contig gap from N315 where applicable or the mean of the estimate from the read pairs otherwise. Then, MGA was used to recompute the MAC, and the .align file produced by MGA was used to carefully scrutinize the MAC by eye and fix the infrequent alignment errors manually. The JH1 and JH9 chromosomes are each estimated to be about 2,900,000 bp long, a figure which includes the contig gaps. The JH1 and JH9 contig gaps are estimated to contain a mere 1.5 and 2.3% respectively of the JH1 and JH9 chromosomal sequence.

Construction of the MACR. We produced a full multi-alignment (referred herein to as the MACR) of the JH1, JH9, and N315 chromosomal sequences and the JH1 and JH9 reads. The program clustalw was used to construct piecewise the MACR from the MAC and the mapping of the JH1 and JH9 reads onto respectively the JH1 and JH9 contigs. The final MACR had a length of about 3,000,000 columns. In the MACR, an indel in a read was assigned the Phred quality value of the previous base in the read. To identify alignment errors in the MACR, we searched for and manually examined columns in which the symbol in a contig sequence disagreed with a symbol in a read with a high Phred quality value. Alignment errors were found to occur at a rate of only about 1:200,000 columns, with no error spanning more than several columns.

Preliminary manual comparison of the JH1, JH9, and N315 chromosomal sequences. Using the .align file outputted by MGA, we carefully examined by eye the entire MAC. Differences between the JH1, JH9, and N315 chromosomal sequences were noted. Considered was the sequence interior to the JH1 and JH9 contigs, which are estimated to contain 98.5% and 97.7% respectively of the JH1 and JH9 chromosomal sequence. There were hundreds of large and small differences between JH1 and N315, including indels >40,000 nucleotides long. The only differences between JH1 and JH9 were isolated ND's, with the exception of some indels 2-20 nucleotides long and clusters of ND's. The indels 2-20 nucleotides long and clusters of ND's were expected to be have been produced by read errors, since they always occurred in the MACR in regions of 1 or 2X coverage with poor read quality. Moreover, the indels ≥ 10 nucleotides always involved poly- A and T sequence at the end of a read, which is usually unreliable.

Identification and subsequent experimental verification of the mutations between the JH1, JH9, and N315 chromosomes. We formulated a Bayesian probabilistic model (BPM) to identify real ND's in the MACR column by column. Thus, bona fide insertions or deletions ≥ 2 nucleotides in length and regions of non-homology that spanned multiple consecutive columns in the MACR were identified column by column. The BPM assumed that the N315 sequence contained no errors and considered the coverage and Phred quality values of the JH1 and JH9 reads. If a real ND could be predicted or ruled out in a column with a probability $P > T =$

$1-1/(3\times10^6)$, then the column was said to be informative. The threshold T was selected so that the expected error rate (both false positive and false negative) was less than one per genome when calling a real ND in an informative column. Such confidence could not be achieved in uninformative columns, which occurred due to poor coverage (mostly 0-1X) and/or read quality. When a real ND was predicted in an informative column, the region containing the column in the MACR was always manually examined. The 10-20 columns with alignment errors in the MACR were checked. Anomalous predictions due to alignment errors were identified and not reported. It was also ensured that no real ND went unreported due to an alignment error. In the JH1 and N315 comparison, 97% of the columns in the MACR were informative, and only the real ND's predicted in the informative columns were reported. Thus, no false ND between N315 and JH1 due to a read error in JH1 is expected to have been reported. In the JH1 and JH9 comparison, 94% of the columns in the MACR were informative. Every predicted real ND in the informative columns was confirmed by PCR sequencing, except for the deletion in the stretch of 14 adenines that could not be sequenced (Table 1 in the main text and Tables S.1 and S.2). Also considered were 10 of the most promising predictions of real ND's in the uninformative columns. All but two were shown to be false by PCR sequencing. One proved to be real, and the other was the deletion in the IS1811 insertion sequence that could not be PCR'd (Tables 1, S.1, and S.2). In total, real ND's occurred at a rate of only 35 in 94% of the columns. Assuming that the NDs occurred at the same rate in uninformative columns, the number of NDs expected to have gone unreported in the 6% of uninformative columns in the MACR can be estimated as $35 / 0.94 \times 0.06 = 2.2$. When a second more rigorous estimate was done that considered the coverage and read quality in the uninformative columns, a number of 1.5 was obtained.

4. Assessment of overlaps of lists of transcriptional changes.

Table M.1. Lists of ORFs considered. The activity of the Agr qorum sensing system is believed to be growth dependent (19, 21). Therefore, it is important to note that the lists JH9/JH1$_0$ and ARLR$_0$ were both determined in mid-exponential phase and that the lists AGR$_0$ and TRAP$_0$ were both determined in post-exponential phase.

Name of list	Description	Ref. for data set	Expected affect of mutation in JH9 on transcriptional regulator
ALL$_0$	all 2588 ORFs on N315 chromosome	(45)	
JH9/JH1$_0$	224 ORFs found to be upregulated or downregulated by ≥ 2-fold in JH9 compared to JH1 (determined in mid-exponential phase)	(41)	
VRAR$_0$	46 ORFs identified to be induced directly or indirectly by VraR	(5)	**Expectation.** There is a nonsynonymous substitution in SA1702, which is in the *vraSR* operon (4). It increases the activity of VraR. **Reason.** In JH9/JH1$_0$, the genes *vraSR* are over-expressed in JH9 compared to JH1.
AGR$_0$	138 ORFs identified to be positively or negatively regulated directly or indirectly by the Agr qorum sensing system (determined in post-exponential phase)	(18)	**Expectation.** Frameshift in *agrC* decreases activity of Agr. **Reason.** Loss of Agr function. See (*21*).
TRAP$_0$	78 ORFs identified to be positively or negatively regulated directly or indirectly by TRAP, a positive regulator of the *agr* locus (determined in post-exponential phase)	(19)	**Expectation.** Frameshift in *agrC* decreases effect of TRAP. **Reason.** Loss of Agr function. See (19, 21).
ARLR$_0$	114 ORFs identified to be positively or negatively regulated directly or indirectly by ArlR, a positive regulator of the *agr* locus (determined in mid-	(20)	**Expectation.** Frameshift in *agrC* decreases effect of ArlR. **Reason.** Loss of Agr function. See (20, 21).

	exponential phase)		
AGR_ALL_0	list of 244 ORFs produced by the union of AGR_0 + $TRAP_0$ + $ARLR_0$		See entries for AGR_0, $TRAP_0$, and $ARLR_0$.
$YYCF_0$	32 ORFs predicted to be directly regulated by YycF	(26)	**Expectation.** Truncation of *yycH* increases activity of YycF. **Reason.** Deletion of *yycH* has been shown to lead to increase in YycF-dependent gene expression in *B. subtilis* (71).

We generated a list of putative operons by grouping any two consecutive tandemly transcribed ORFs on the N315 chromosome into the same putative operon if the two ORFs were separated by less than 50 bp of intergenic sequence. For each list X_0 of ORF's in Table M.1, we generated a new list X_{50} consisting of the putative operons each with at least one ORF in X_0.

For $n = 0$ or 50, we assessed the overlap between the list X_n controlled by a transcription factor and the list $JH9/JH1_n$ as follows:

a) Each locus (ORF when $n = 0$ and operon when $n = 50$) in X_n that is positively (negatively) regulated by the factor was assigned a +1 (-1) if the mutation in JH9 was expected to increase the activity of the factor. Alternatively, each locus in X_n that is positively (negatively) regulated by the factor was assigned a -1 (+1) if the mutation in JH9 was expected to decrease the activity of the factor.

b) Each locus in $JH9/JH1_n$ that is upregulated (downregulated) in JH9 compared to JH1 was assigned a +1 (-1).

c) We determined the overlap between X_n and $JH9/JH1_n$, defined as the number N of loci that not only appear in both X_n and $JH9/JH1_n$ but are assigned 1's with the same sign in the two lists.

d) Under a null model that assumed the two lists X_n and $JH9/JH1_n$ were picked independently, we computed the expected overlap μ of the two lists:

$$\mu = \frac{\left|\mathrm{JH9/JH1}_n\right|}{\left|\mathrm{ALL_n}\right|}\frac{1}{2}\left|X_n\right| \tag{1}$$

where $|\bullet|$ denotes the size of the list.

e) A P-score P for the observed overlap N was computed from the Poisson distribution:

$$P = \sum_{k \geq N} \frac{\mathrm{e}^{-\mu}\mu^k}{k!}. \tag{2}$$

The P-scores are summarized in Table M.2.

The P-scores for the lists of ORFs were considered less reliable than for the lists of putative operons. When regarding ORFs as independent, there is a danger of overestimating the significance of overlap when a large operon consisting of many ORFs happens to make it into both lists. Note that the lists for Agr do not have good P-scores even though the lists for Trap and AlrR do. This may have to do with the fact that the list for Agr were determined in post-exponential phase whereas the lists for Trap and AlrR and the transcriptome profile of JH9 versus JH1 were determined in mid-exponential phase.

Table M.2. The P-score of the overlap of X_n with JH9/JH1$_n$.
The highlighted P-scores are those reported in the main paper.

X_n	P-score	X_n	P-score
$VRAR_0$	10^{-13}	$VRAR_{50}$	10^{-11}
AGR_0	$10^{-0.7}$	AGR_{50}	$10^{-0.5}$
$TRAP_0$	10^{-4}	$TRAP_{50}$	10^{-4}
$ARLR_0$	10^{-3}	$ARLR_{50}$	10^{-2}
AGR_ALL_0	10^{-3}	AGR_ALL_{50}	10^{-2}
$YYCF_0$	10^{-2}	$YYCF_{50}$	10^{-2}

5. Available for download or upon request

Table M.3. Access to data and programs.

Item	Section first discussed	Reference	How to obtain
Sequence data			
raw JH1 and JH9 reads with base specific Phred quality values	7		NCBI's trace archive[a].
trimmed JH1 and JH9 reads with base specific Phred quality values	7		Get a copy of data.tar.gz[b].
complete JH1 and JH9 plasmid sequences	8		
full multi-alignment of JH1, JH9, and N315 chromosomal sequences and trimmed JH1 and JH9 reads with site specific Phred quality values (i.e. MACR)	10		
Assembler			
recent version wgs-assembler-3.10 of the Celera assembler	7	(64)	[c]
Alignment programs			
clustalw, version 1.81	6	(66)	See references.
dialign, version 2.2		(68)	
MGA, March 18, 2003 release		(65)	
q-gram filters	7	(70)	
blast programs	7	(72)	
Programs written for this study			
PERL script myalign.pl called by MGA to invoke clustaw and dialign	6		Get a copy of data.tar.gz[b].
Previously sequenced genomes and lists of open reading frames (i.e. predicted genes)			
S. aureus strains N315, Mu50, MW2, COL, 476, and 252; *E. coli K12*; human.	7		NCBI's genome database[d]

a. Go to the webpage:
http://www.ncbi.nlm.nih.gov/Traces/trace.cgi?cmd=stat&f=xml_list_species&m=obtain&s=species.

b. The sequence data and PERL script myalign.pl have been combined into a single compressed archive file data.tar.gz using the Unix tools tar and then gzip. A copy of data.tar.gz can be requested by emailing the first author Michael Mwangi at:
mwangi@morel.rockefeller.edu.

c. The current webpage for the wgs-assembler is:
http://sourceforge.net/projects/wgs-assembler/.

d. Go to the webpage:
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB=genome.

6. Multi-alignment programs used

We used the following multi-alignment programs:

a) clustalw, version 1.81 (66)
b) dialign, version 2.2 (68)
c) MGA, March 18, 2003 release (65)

Suitable for multi-aligning numerous small sequences at a time, clustalw and dialign were used frequently – always though on no more than several dozen sequences totaling no more than a few 100,000 nucleotides in length. When aligning *S. aureus* sequences that differed by isolated ND's, clustalw produced better alignments than dialign, particularly for the more dissimilar sequences. When aligning *S. aureus* sequences that were nearly identical if not for large indels due to strain specific genomic islands or gaps in sequence, dialign produced better alignments than clustalw. Unless stated otherwise, clustalw and dialign were invoked as follows:

clustalw -type=DNA -dnamatrix=IUB -output=GDE -outorder=input -pwgapext=0 -gapext=0 -infile=[enter here input file name]

dialign2-2 -n -strict -thr 5 -fa [enter here input file name]

Capable of globally multi-aligning several closely related bacterial chromosomes at a time, MGA was used to produce multi-alignments of large segments and/or whole chromosomes. To align regions between maximal exact matches, MGA calls external programs via a user specified script. A PERL script myalign.pl was written so that MGA would call clustalw when the regions differed in length by less than 10 nucleotides and dialign otherwise. The program MGA was invoked as follows:

mkvtree -dna -lcp -suf -tis -indexname [enter here index name] -db [enter here list of files each containing one of the sequences to be multi-aligned]

mga.32seqs -v -l 1000 -always -gl 100000 -msascript myalign.pl -alignedseqs -gap -width 100 -o [enter here prefix of output files] [enter here index name specified in call to mkvtree above]

The .align file produced by MGA proved to be especially useful since the file is a concise human readable list of all the differences between sequences in the MGA muli-alignment. The file therefore permitted the careful inspection by eye of large machine generated global muli-alignments.

7. Sequencing and *de novo* assembly of the JH1 and JH9 genomes and mapping of the JH1 and JH9 reads onto respectively the JH1 and JH9 contigs

7.1. Whole genome shotgun sequencing of JH1 and JH9

For each of JH1 and JH9, the whole genome shotgun sequencing (62, 63) was carried out by the Joint Genomes Institute (JGI) (69) as follows:

a) Genomic DNA was sheered into random fragments, size selected, and cloned into an appropriate vector to produce three different sized libraries of clones:
   Genomic DNA was randomly sheared using a hydroshear device (Genemachines, San Carlos, CA), and the fragments were blunt-end repaired with T4 polymerase and Klenow fragment. Fragments were size selected by agarose gel electrophoresis, purified from the gel (Qiaquick, Qiagen Corporation, Valencia, CA), and ligated into pUC18 (small inserts), pMCL200 (medium inserts), or pCC1Fos (large inserts) (Epicentre, Madison, WI). Ligations were transformed into *E. coli* DH10B cells, and colonies were picked into 384-well plates containing LB and glycerol.
b) The ends of the clones were sequenced to generate over 40,000 single reads with a mean length between 500 to 1000 nucleotides:
   DNA for sequencing was produced by rolling circle amplification (Templiphi, GE Healthcare, Piscataway, NJ) or Sprintprep (Agencourt, Beverly MA) magnetic bead DNA purification. Subclone inserts were sequenced from both ends using universal primers and ET (GE Heathsciencies, Piscataway, NJ) or Big Dye (ABI, Foster City, CA) terminator chemistry.
c) Based on the trace data, each base call in each read was assigned its own Phred quality value Q, which was always an integer from 0 to 60 such that $10^{-Q/10}$ is the probability the base call is incorrect (67).
d) In most cases, both ends of a clone were sequenced, generating two single reads that formed a read pair. Two reads in a read pair are said to be mates.

The sequencing protocols of the JGI are described in detail in (73) and at http://www.jgi.doe.gov/sequencing/protocols/index.html.

7.2. Trimming of the reads

For each of JH1 and JH9:
a) The raw unprocessed reads were trimmed for both vector and quality using a specialized trimming pipeline.
b) The sum N of the lengths of the trimmed reads was between 26,000,000 to 29,000,000 nucleotides.
c) Previously, it was established (44) by pulse field gel electrophoresis that the size L of the JH genome is similar to the size of the N315 genome, known to contain about 3,000,000 bp. Hence, the mean coverage N/L of the trimmed reads is estimated to be between 8.5 to 9.5X.

7.3. Sizes of the libraries of clones and definition of a read pair's mean and standard deviation

Since MLST typing had suggested that the JH lineage differs from the sequenced *S. aureus* strain N315 by a point mutation rate as small as 1:3000 bp (44), the sizes of the three JH1 and three JH9 libraries of clones in Section 7.1 were determined using N315. For each library, the trimmed reads were mapped to the N315 genomic sequence using the q-gram technique (70), and the distance between the outermost ends of the trimmed reads in a read pair was found to be an approximately normally distributed random variable. For JH1, the distance had a mean of 3, 6, or 35 kbp and a standard deviation of 0.3, 0.5, or 3 kbp respectively, depending on which of the three libraries was considered. For JH9, the mean was 3, 6, or 36 kbp, and the standard deviation was 0.3, 0.5, or 4 kbp respectively.

For convenience, we speak of a read pair's mean and standard deviation. By this, we are actually referring to the mean (3, 6, or 35-36 kbp) and the corresponding standard deviation (0.3, 0.5, or 3-4 kbp respectively) computed above for the library of clones containing the read pair.

7.4. *De novo* assembly of the trimmed reads into contigs and scaffolds and verification of the sizes of the libraries of clones

For each of JH1 and JH9, the *de novo* assembly of the trimmed reads was done using a recent version wgs-assembler-3.10 of the Celera assembler, the first version of which was developed by a team led by the coauthor Myers (64). Invoked using standard parameter values, the assembler took as input: a list specifying the library and mate of each read (from Section 7.1), the trimmed reads (Section 7.2), and estimates of the sizes of the libraries of clones (Section 7.3). The assembler assembled the reads into contigs and grouped the contigs into scaffolds. A contig is a continuous segment of known sequence representing the consensus of overlapping staggered reads. A scaffold is a set of contigs ordered by position and all orientated to represent the same strand (Watson or Crick), plus estimates of the sizes of all the intra-scaffold contig gaps of unknown sequence between consecutive contigs. In a scaffold, the assembler can infer the contigs' order and strands and can estimate the sizes of the contig gaps using read pairs bridging contigs and the estimates of the sizes of the libraries of clones. The assembler's estimate of the size of a contig gap is a normally distributed random variable with a specified mean and standard deviation. As a consistency check, the assembler re-estimates the sizes of the libraries of clones using read pairs it maps to the same contig. The sizes computed in Section 7.3 and the assembler's re-estimates agreed.

7.5. First unambiguous mapping of trimmed reads to contigs

Although rare, the Celera assembler can fail to group two contigs into a single scaffold even though the contigs are connected by read pairs. In the case of highly similar but non-identical repetitive regions, the assembler can also incorrectly map a read to the wrong region. For each of JH1 and JH9, we therefore felt it was prudent to also work with an unambiguous mapping of the trimmed reads to contigs, which was generated as follows:

a) The reads were locally aligned to the contigs using the q-gram technique (70).
b) For each read: All matches that had >97% nucleotide identity along their length were found. Then, matches that were <80% of the length of the longest match were eliminated. Afterwards, the best remaining match was identified, defined as the match with the smallest error rate. Finally, matches with an error rate >2 times the error rate of the best match were eliminated.
c) For each remaining match of each read: If for the match of the read it could be ruled out that there is a match of the read's mate with a suitable orientation so that the two matches are a proper distance apart from each other on the genome, then the match of the read was eliminated. To be a proper distance apart, the two matches have to be separated on the genome by a distance that agrees with the read pair's mean to within three of the read pair's standard deviations (see Sections 7.3 and 7.4).
d) Reads still with multiple matches were eliminated.

7.6. Elimination of contaminant sequence by comparison with previously sequenced *S. aureus* strains

Contigs arising from contaimnant DNA were eliminated. Firstly, all the contigs were blasted against the completely sequenced genomes of the *S. aureus* strains N315, Mu50, MW2, COL, 476, and 252 using BLASTN with no filter (72). A contig with no homology (E-score < 10^{-5}) to any of the *S. aureus* genomes was discarded, provided that the contig was not associated with another contig having homology to at least one of the *S. aureus* genomes. Two contigs were said to be associated with one another only if they were grouped together in a scaffold in Section 7.4 or connected by a read pair in the unambiguous mapping in Section 7.5. Thus, seemingly foreign sequence not in the previously sequenced *S. aureus* strains was eliminated only when there was no evidence connecting it to known *S. aureus* sequence. For each of JH1 and JH9, <2% of the contig sequence was gotten rid of. All the contigs that were discarded were 100-5000 nucleotides in length. About 10% of the eliminated sequence had >99% nucleotide identity to the sequenced *E. coli* K12 or human genomes. In the end, 62 JH1 and 79 JH9 contigs remained.

8. Determination of the JH1 and JH9 plasmid and chromosomal sequence and ordering of the JH1 and JH9 contigs using N315

8.1. Preliminary comparison of JH1, N315, and Mu50

Previously, all the JH isolates JH1-JH15 were found to have the MLST 1-4-1-4-12-1-28, and both of the sequenced *S. aureus* strains N315 and Mu50 were found to have the MLST 1-4-1-4-12-1-10 (44). To determine whether JH1 is more closely related to N315 or Mu50, we used MGA to multi-align several randomly chosen large JH1 contigs with the N315 and Mu50 genomic sequences. By inspecting the .align file outputted by MGA (see Section 6), we found that JH1 agreed in about three out of every four cases with N315 rather than Mu50 when we examined over 50 isolated point mutations chromosome-wide between N315 and Mu50.

8.2. Determination of the plasmid and chromosomal sequence and ordering of the contigs using N315

To order the scaffolds in Section 7.4, we blasted the scaffolds against the N315 genomic sequence using BLASTN with no filter (72). The N315 genome consists of a circular 24,653 bp plasmid and a circular 2,814,816 bp chromosome.

One scaffold in each of JH1 and JH9 exhibited significant homology to the N315 plasmid but little homology to the N315 chromosome. Each of these two scaffolds consisted of only one contig slightly greater than 30,000 nucleotides. It became clear that each of these two contigs represented the complete sequence of a circular plasmid for the following reasons. Each contig exhibited >99% nucleotide identity to the N315 plasmid over a 15,000 nucleotide region. When each contig was inspected, it was found that the first roughly 1000 nucleotides at one end of the contig and the last 1000 nucleotides at the other end were identical, suggesting the two ends were

not distinct. For each contig, many read pairs in the unambiguous mapping in Section 7.5 mapped to the contig such that the distance between a read and its mate on the contig could only be reconciled with the read pair's mean and standard deviation (see Sections 7.3 and 7.4) if the contig represented a circular DNA molecule with a length of about 30,000 bp.

It was determined that the remaining JH1 and JH9 scaffolds represented chromosomal sequence. The following applies to each of JH1 and JH9: Although dozens of large insertions and deletions 1000-50,000 bp were observed between the JH and N315 chromosomes, the vast majority of the JH scaffolds exhibited enough homology over at least some of their contigs' lengths that they could be mapped unambiguously to the N315 chromosome. These JH scaffolds could be ordered by position using N315. To make all scaffolds represent the Watson strand, scaffolds matching the Crick strand of the N315 chromosome were re-orientated by reverse complementation. The several small JH scaffolds that exhibited little or no homology to the N315 chromosome each consisted of only one contig < 10,000 bp, representing novel chromosomal sequence in the JH lineage. The positions and suitable orientations of these small scaffolds could be inferred using read pairs in the unambiguous mapping in Section 7.5 that connected the small scaffolds to the scaffolds already ordered using N315. When the ordering of the contigs was complete, it could be seen that the JH chromosome was syntenous over its entire length with the N315 chromosome, except for several transpositions involving elements < 10,000 bp.

8.3. Second unambiguous mapping of trimmed reads to contigs

Since some of the scaffolds were re-orientated in Section 8.2, the unambiguous mapping of the trimmed reads to contigs in Section 7.5 was re-done just as before.

9. Estimation of the sizes of the JH1 and JH9 contig gaps, the ruling out of large differences between JH1 and JH9 in the contig gaps, and the editing of the MAC

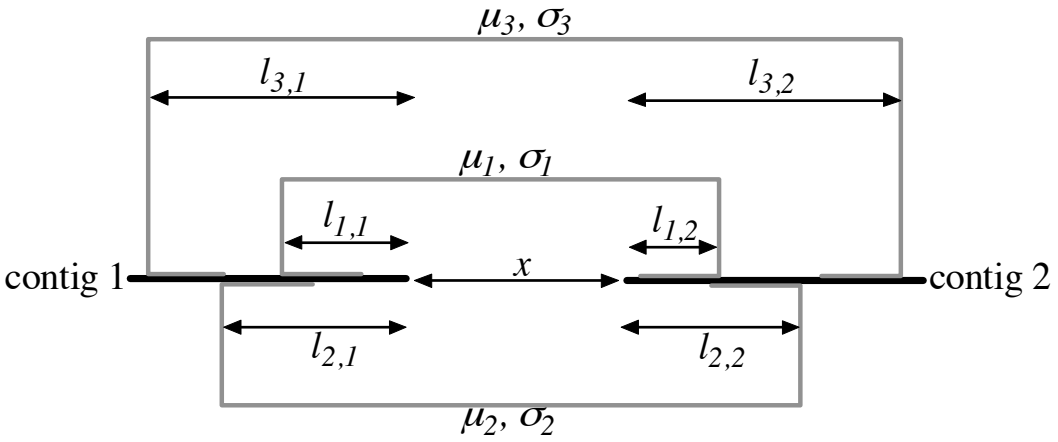
9.1. Our estimates of the sizes of the contig gaps from read pairs

In Section 7.4, it was stated that assembler estimated the sizes of all intra-scaffold contig gaps from read pairs. Independently of the assembler, we estimated the sizes of contig gaps using read pairs in the unambiguous mapping produced in Section 7.5. To estimate the size of a contig gap, we used read pairs that span the gap and connect the two consecutive contigs to the gap's immediate left and right. When such read pairs were not available, an attempt was made to use read pairs that span the gap and connect non-consecutive contigs. Regardless, the estimation was done using the same Bayesian approach. As an illustration, a contig gap is depicted in Figure M.1. We would estimate the size of the contig gap as follows:

a) In the absence of prior information, we use an uninformative prior $\mathrm{P}(x)$ over x:

$$\mathrm{P}(x) = \begin{cases} 1/(L+U+1), & x = -L, -L+1, ..., U \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

Figure M.1: Estimation of the size x in nucleotides of a contig gap. In this case, there are $N > 0$ read pairs that span the gap and connect the two consecutive contigs 1 and 2 to the gap's immediate left and right respectively. For simplicity, only three read pairs are shown. For the i^{th} read pair, several values are listed: μ_i and σ_i denote the read pair's mean and standard deviation (see Sections 7.3 and 7.4), and $l_{i,j}$ denotes the distance in nucleotides from the outermost end of the read mapping to the contig $j = 1$ or 2 to the furthermost end of the contig.



for $L,U \geq 0$. That is, x occurs with equal probability anywhere in the interval $[-L,U]$ and with zero probability elsewhere. Here, the values of L and U are not particularly significant. We work in the limit of large L and U. Thus, L and U serve only to indicate that little is known a priori about the value of x.

b) Since it is always the case that $\sigma_i >> 1$, the probability $\mathrm{P}(D \mid x)$ of observing the particular configuration D of read pairs given x is:

$$\mathrm{P}(D \mid x) = \prod_{i=1}^{N} \left\{ \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left[-\frac{\left(l_{i,1} + x + l_{i,2} - \mu_i\right)^2}{2\sigma_i^2} \right] \bullet 1 \right\}. \tag{4}$$

c) By Bayes' theorem, the posterior probability of x given the configuration D of read pairs is:

$$\mathrm{P}(x \mid D) = \frac{\mathrm{P}(D \mid x)\mathrm{P}(x)}{\sum_{x=-\infty}^{\infty} \mathrm{P}(D \mid x)\mathrm{P}(x)} = \frac{\exp\left[-\frac{\left(x-\mu\right)^2}{2\sigma^2} \right]}{\sum_{x=-L}^{U} \exp\left[-\frac{\left(x-\mu\right)^2}{2\sigma^2} \right]} \tag{5}$$

for $-L \leq x \leq U$ where

$$\mu = \sum_{i=1}^{N} \frac{1/\sigma_i^2}{\sum_{j=1}^{N} 1/\sigma_j^2} \left(u_i - l_{i,1} - l_{i,2} \right) \tag{6}$$

and

$$\sigma = \sqrt{\frac{1}{\sum_{i=1}^{N} 1/\sigma_i^2}} \,. \tag{7}$$

Since we are working in the limit of large L and U and it is always observed that $\sigma \gg 1$, the following approximation can be made:

$$\sum_{x=-L}^{U} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2} \right] = \int_{-\infty}^{\infty} \mathrm{d}x \exp\left[-\frac{(x-\mu)^2}{2\sigma^2} \right] = \sigma\sqrt{2\pi} \,. \tag{8}$$

Hence,

$$\mathrm{P}(x \mid D) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2} \right]. \tag{9}$$

Thus, the estimate of the size x of the contig gap is a normally distributed random variable with a mean μ and standard deviation σ.

9.2. Estimating the sizes of the JH1 and JH9 contig gaps and ruling out large differences between JH1 and JH9 in the contig gaps

By inspecting the .align file produced by MGA (see Section 6), we were able to very carefully examine by eye the entire MAC. Save for the few JH1 and JH9 contig gaps in new sequence specific to the JH lineage, it was possible to infer the sizes of all gaps from the corresponding sequence in N315. Thus, sometimes as many as three separate estimates of the size of a contig gap were available – one inferred from N315, another computed by the Celera assembler from read pairs (see Section 7.4), and yet another computed by us from read pairs (Section 9.1). Ultimately, at least one estimate was available for the size of every contig gap. For lists of all the estimates, see Tables M.4 and M.5.

In several cases, contigs were found to overlap. To ensure that the apparent overlap was genuine and not the result of a tandem duplication, the read pairs spanning the putative region of overlap were carefully examined and in one case restriction digest and Southern analysis was even done (see the comments in Tables M.4 and M.5).

When there was an indication of a large difference between JH1 and JH9 in a contig gap, the unknown sequence in the gap was determined by PCR sequencing. In each case, JH1 and JH9 were found to be identical (see the comments in Tables M.4 and M.5).

Table M.4. JH1 contig gaps.

First contig[a]		Subsequent contig[a]		Estimates of the size of the contig gap in nucleotides[b]					Comments
Unique identifier of contig	Length in nucleotides[b]	Unique identifier of contig	Length in nucleotides[b]	N315[c]	Assembler[d]		Us[e]		
					mean	STD	mean	STD	
1.1	404980	1.2	34200	149	541	959	497	858	
1.2	34200	1.3.1	211627	483	258	127	249	123	
1.3.1	211627	1.3.2	75805		3010	520			This gap in JH1 is in a phage-like genomic island > 40,000 nucleotides specific to JH1 and JH9.
1.3.2	75805	1.4	10562	1797	1256	1119	1541	1357	
1.4	10562	1.5	1551	532	878	302	887	297	
1.5	1551	1.6	42793		1530	1598			This gap in JH1 overlaps an IS1811 transposon insertion that is in N315 but not JH9. Because the assembler's estimate[d] of the size of the gap has a large uncertainty ± 1598 nucleotides, PCR sequencing was done to show that the sequence in the gap in JH1 does not contain the IS1811 insertion in N315 and is instead identical to JH9.
1.6	42793	1.7	6739	75	-20	105	-35	100	
1.7	6739	1.8	3783	953	-20	2723			
1.8	3783	1.9	31687	349	-20	1073	3042	1919	
1.9	31687	1.10	23352	46	166	155	73	135	
1.10	23352	1.11	10895	281	1260	295	1360	297	The assembler's and our estimates[d, e] of the size of the gap in JH1 are respectively 1260 ± 295 and 1360 ± 297 nucleotides. However, the corresponding regions in N315 and JH9 are both only 281 nucleotides long. The fact that the estimates of the size of the gap in JH1 differ by more than three standard deviations from the N315 and JH9 lengths suggests that there is an insertion specific to JH1. Nevertheless, PCR sequencing confirmed that the sequence in the gap in JH1 contained no insertion and is instead identical to N315 and JH9.
1.11	10895	1.12	1548	59	378	455	441	515	
1.12	1548	2.1	1660	285					*The assembler's estimate[d] of the size of the gap between contigs 1.12 and 1.13 in JH1 is 3557 ± 453 nucleotides. Thus, the size of the gap between 2.1 and 1.13 in JH1 is about 3557−285−1660 = 1612 nucleotides. In all *S. aureus* strains sequenced thus far, there are several occurrences in the chromosome of a few thousand nucleotides long segment coding for tRNA's and rRNA's. One occurrence is found in this region in N315, JH1, and JH9. However, the segment appears to be longer in JH1 and JH9, perhaps due to a tandem duplication. The additional sequence includes parts of contigs 1.13 and 1.14. That the additional sequence is not an artifact of the assembly is supported by reads spanning the novel juncture between the region common to N315, JH1, and JH9 and the additional region specific to JH1 and JH9. There is no evidence of a difference between JH1 and JH9.
2.1	1660	1.13	1540	*See comments.					
1.13	1540	1.14	1963		501	215	512	209	
1.14	1963	1.15	152230	1409	440	332	1393	255	
1.15	152230	1.16	6428	110	-20	162	-90	117	
1.16	6428	1.17	972		452	215	455	209	This gap in JH1 is in an IS1181 transposon insertion specific to JH1 and JH9.
1.17	972	1.18	229181	372	1280	430			
1.18	229181	1.19	1299	1651	1872	357	1964	364	
1.19	1299	1.20	59738	1329	-20	1734			
1.20	59738	1.21	78318	301	39	127	35	124	
1.21	78318	3.1	1695	0					
3.1	1695	3.2	1325	244	484	215	493	209	The N315 estimate[c] of the size of the gap between contigs 3.1 and 4.1 in JH1 is 2350 nucleotides, which agrees to within two standard deviations with our estimate[e] of 3036 ± 515 nucleotides.
3.2	1325	4.1	1562	781					
4.1	1562	5.1.1	18211	2226					
5.1.1	18211	5.1.2	46493	0	-20	456			

5.1.2	46493	5.2	26371	977	306	502	342	481	
5.2	26371	6.1	57445	4730					
6.1	57445	7.1	1160	0					The N315 estimate[c] of the size of the gap between contigs 6.1 and 6.2 in JH1 is 3746 nucleotides, which agrees to within one standard deviation with the assembler's estimate[d] of 4081 ± 2350 nucleotides.
7.1	1160	8.1	1567	916			1032	296	
8.1	1567	6.2	55053	75					
6.2	55053	9.1	11552	1796			-2020	3324	
9.1	11552	9.2	70828	314	-19	120	-19	117	
9.2	70828	9.3	16186	103	324	140	324	137	
9.3	16186	9.4	7955	113	112	131	112	128	
9.4	7955	9.5	1187	497	681	242	897	296	
9.5	1187	9.6	11020	160	122	224	267	257	
9.6	11020	9.7	11802	563	-20	403	-169	364	
9.7	11802	9.8	300702	272	472	198	456	194	
9.8	300702	9.9	2868	289	295	163	251	170	
9.9	2868	9.10	19351	292	347	220	229	256	
9.10	19351	9.11	2349	900	917	303	926	296	
9.11	2349	10.1	74807	310			248	257	
10.1	74807	10.2	161502	185	-20	1345	-201	1002	
10.2	161502	10.3	87353	451	317	147	408	109	
10.3	87353	10.4	22559	55	210	158	207	154	
10.4	22559	11.1	9097	203			112	104	
11.1	9097	12.1	13517	0					
12.1	13517	12.2	16205	78	216	165	206	162	
12.2	16205	12.3	24143	208	191	173	183	170	
12.3	24143	12.4	126471	446	1140	807	816	441	
12.4	126471	12.5	56414	48	-20	172	-57	153	
12.5	56414	13.1	61722	6595					
13.1	61722	13.2	63541	598	227	1175	363	1108	
13.2	63541	14.1	1075	350					The N315 estimate[c] of the size of the gap between contigs 13.2 and 14.2 in JH1 is 1842 nucleotides, which agrees to within two standard deviations with our estimate[e] of 2584 ± 515 nucleotides.
14.1	1075	14.2	27500	416	338	264	377	162	
14.2	27500	14.3	40576	147	96	215	115	208	
14.3	40576	14.4	9376	203	357	141	369	137	
14.4	9376	15.1	7508	408					
15.1	7508	1.1	404980	0					

a. "First contig" and "Subsequent contig" are the two consecutive contigs directly flanking the contig gap. ***b.*** In several cases, contigs were found to overlap. The redundant duplicate sequence was eliminated, and the lengths of the contigs and estimates of the sizes of the contig gaps had to be updated accordingly. This table lists not the initial but the adjusted lengths and estimates. ***c.*** The N315 inferred size of the contig gap. ***d.*** The Celera assembler's estimate of the size of the intra-scaffold contig gap computed from read pairs (see Section 7.4). ***e.*** Our estimate of the size of the contig gap computed from read pairs (Section 9.1).

Table M.5. JH9 contig gaps.

First contig[a]		Subsequent contig[a]		Estimates of the size of the contig gap in nucleotides[b]					Comments
Unique identifier of contig	Length in nucleotides[b]	Unique identifier of contig	Length in nucleotides[b]	N315[c]	Assembler[d]		Us[e]		
					mean	STD	mean	STD	
1.1	269456	1.2	115707	633	531	1676	104	1417	It was observed that the first 7200 nucleotides of contig 1.1 and the last 7200 nucleotides of 23.3 in JH9 were identical. The duplicate sequence could have arisen two ways. Firstly, it may be an artifact of the assembly, arising because the assembler inexplicably failed to merge two highly overlapping contigs. Alternatively, the duplicate sequence may be the result of a genuine tandem duplication specific to JH9. Unfortunately, it was not possible to distinguish between the two possibilities by examining the reads or read pairs. Thus, restriction digest followed by Southern analysis was done to determine the length of the region. It was clear from the results that the duplicative sequence was indeed an artifact of the assembly and not the result of

									a tandem duplication. Thus, the redundant duplicate sequence in 1.1 was eliminated. The reported length of 1.1 and the estimates of the sizes of the contig gap between 1.1 and 23.3 were adjusted accordingly.
1.2	115707	1.3	128896	10	-20	252	-269	154	
1.3	128896	1.4	3393		6726	1890	3922	386	This gap in JH9 is in a phage-like genomic island > 40,000 nucleotides specific to JH1 and JH9.
1.4	3393	1.5	44822	417	113	248	203	170	
1.5	44822	1.6	27344	236	654	376	774	372	
1.6	27344	1.7	13533	624	451	522	651	534	
1.7	13533	1.8	1155	318	343	390	345	386	
1.8	1155	1.9	48602	755	486	1601			
1.9	48602	1.10	24555	4162	1733	2164	1729	2164	
1.10	24555	1.11	1524	390	466	388	404	386	
1.11	1524	1.12	16045	1782	1031	388	1937	273	
1.12	16045	1.13	23448	959	775	2232			
1.13	23448	2.1	1469	676			379	263	*The assembler's estimate[d] of the size of the gap between contigs 1.13 and 1.14 in JH9 is 4059 ± 378 nucleotides. Thus, the size of the gap between 2.1 and 1.14 in JH9 is about 4059−676−1469 = 1914 nucleotides. In all *S. aureus* strains sequenced thus far, there are several occurrences in the chromosome of a few thousand nucleotides long segment coding for tRNA's and rRNA's. One occurrence is found in this region in N315, JH1, and JH9. However, the segment appears to be longer in JH1 and JH9, perhaps due to a tandem duplication. The additional sequence includes parts of contigs 1.14 and 1.15. That the additional sequence is not an artifact of the assembly is supported by reads spanning the novel juncture between the region common to N315, JH1, and JH9 and the additional region specific to JH1 and JH9. There is no evidence of a difference between JH1 and JH9.
2.1	1469	1.14	2050	*See comments.					
1.14	2050	1.15	1406		120	237	-131	263	
1.15	1406	1.16	22847	1679	1444	218	1428	193	
1.16	22847	1.17	12619	39	441	1699			
1.17	12619	1.18	123025	292	220	274	133	188	
1.18	123025	1.19	86421		776	167	767	162	This gap in JH9 is in an IS1181 transposon insertion specific to JH1 and JH9.
1.19	86421	1.20	18297	223	-20	240	-69	222	
1.20	18297	1.21	31885	241	212	115	220	112	
1.21	31885	1.22	5386	101	-20	150	5	138	
1.22	5386	4.1	88657	94					
4.1	88657	4.2	77395	2213	2158	1874	1615	1676	
4.2	77395	4.3	57649	277	327	159	315	151	
4.3	57649	4.4	3671	18	134	159	97	133	
4.4	3671	5.1	67985	9083					
5.1	67985	5.2	8869	211	244	113	200	110	
5.2	8869	5.3	15073	160	241	159	242	157	
5.3	15073	6.1	1531	1527					
6.1	1531	7.1	4797	380					
7.1	4797	7.2	27160	109	-20	133	-105	99	
7.2	27160	7.3	25357	47	-20	130	-28	111	
7.3	25357	8.1	1334	791					The N315 estimate[c] of the size of the gap between contigs 7.3 and 8.3 in JH9 is 5776 nucleotides, which agrees to within one standard deviation with our estimate[e] of 8661 ± 3749 nucleotides.
8.1	1334	8.2	3265	52	186	161	192	151	
8.2	3265	8.3	52445	335	195	160	191	153	
8.3	52445	8.4	3708	4323	4990	2665			
8.4	3708	8.5	13281	140	192	276	426	153	
8.5	13281	9.1	48596	565					

9.1	48596	9.2	16592	1308	1391	208	1388	206	
9.2	16592	9.3	9190	100	-20	141	-67	118	
9.3	9190	9.4	4092	491	702	225	704	222	
9.4	4092	9.5	3501	1742	1560	276	1562	273	
9.5	3501	10.1	3604	943			1384	546	
10.1	3604	10.2	10832	0	81	318	83	315	
10.2	10832	11.1	10349	886					
11.1	10349	11.2	62638	303	448	184	8	244	
11.2	62638	11.3	238542	285	271	152	270	150	
11.3	238542	11.4	870	415	588	196	597	186	
11.4	870	11.5	47651	637	802	276	818	263	
11.5	47651	11.6	1422	622	832	226	833	217	
11.6	1422	12.1	166625	2655					
12.1	166625	12.2	20228	202	246	276	244	273	
12.2	20228	14.1	77858	542					
14.1	77858	14.2	5300	21	110	120	109	117	
14.2	5300	14.3	49117	18	-8	108	-8	105	
14.3	49117	16.1	3682	5343			6272	3749	
16.1	3682	17.1	50904				2378	546	
17.1	50904	17.2	30014	828	3620	1874	3303	1676	
17.2	30014	17.3	88036	876	532	382	530	376	
17.3	88036	17.4	1208	430	753	205	566	175	
17.4	1208	17.5	48797	84	285	206	435	236	
17.5	48797	17.6	9441	189	102	546	100	540	
17.6	9441	18.1	1071	709			403	546	The N315 estimate[c] of the size of the gap between contigs 17.6 and 17.7 in JH9 is 1786 nucleotides, which agrees to within one standard deviation with the assembler's estimate[d] of 1692 ± 153 nucleotides.
18.1	1071	17.7	3784	6					
17.7	3784	19.1	62052	4873					
19.1	62052	20.1	1500	0					
20.1	1500	19.2	62027	120					
19.2	62027	21.1	28318	94					
21.1	28318	21.2	34136	367	-20	568	-142	524	
21.2	34136	21.3	5615	912	845	195	843	192	
21.3	5615	21.4	9327	0	-20	568	-9	244	
21.4	9327	22.1	1861	800					
22.1	1861	22.2	6174	2	-20	165	-95	132	The sequence at the end of a read can be unreliable. When the sequence is unreliable, it is frequently clipped off in the trimming stage, but sometimes, it is retained, because of deceptively high quality values. Even when the sequence though unreliable is kept, it can often be identified, since it is typically poly- A or T sequence. When it was observed that the first 14 nucleotides tctagaggatccca in contig 22.2 in JH9 are neither in N315 nor JH1, it was believed that the sequence was spurious, since it had a coverage of only 1X and was assembled from the end of a read. However, tctagaggatccca is not poly- A or T sequence, so it was confirmed by PCR sequencing that the sequence is indeed spurious.
22.2	6174	23.1	11440	103					
23.1	11440	23.2	42258	121	191	160	187	153	
23.2	42258	23.3	81629	0	-20	286	-64	267	
23.3	81629	1.1	269456	0			160	131	

a. "First contig" and "Subsequent contig" are the two consecutive contigs directly flanking the contig gap. ***b.*** In several cases, contigs were found to overlap. The redundant duplicate sequence was eliminated, and the lengths of the contigs and estimates of the sizes of the contig gaps had to be updated accordingly. This table lists not the initial but the adjusted lengths and estimates. ***c.*** The N315 inferred size of the contig gap. ***d.*** The

Celera assembler's estimate of the size of the intra-scaffold contig gap computed from read pairs (see Section 7.4). ***e.*** Our estimate of the size of the contig gap computed from read pairs (Section 9.1).

9.3. The editing of the MAC

The estimates of the sizes of the JH1 and JH9 contig gaps were used to edit the MAC. Each X marking a JH1 or JH9 contig gap was replaced by a string of N's with a length equal to:

a) The estimate of the size of the contig gap from N315 where applicable.
b) Or the estimate computed by the assembler from read pairs as a second resort.
c) Or the estimate computed by us from read pairs as a last resort.

Then, MGA was used to recompute the MAC, and the .align file produced by MGA (see Section 6) was used to carefully scrutinize the MAC by eye and fix the infrequent alignment errors manually.

9.4. Third unambiguous mapping of trimmed reads to contigs

Since some of the contigs were edited in Section 9.2, the unambiguous mapping of the trimmed reads to contigs in Section 8.3 was re-done just as before.

10. Construction of the MACR

We already had the MAC – that is the multi-alignment of the JH1, JH9, and N315 chromosomal sequences, consisting of the ordered JH1 and JH9 contig sequences assembled from and therefore a consensus of the reads. We already had a mapping of the trimmed JH1 and JH9 reads onto respectively the JH1 and JH9 contig sequences (see Section 9.4).

We therefore were able to piecewise construct the MACR. We considered a window of W columns in the MAC plus in this window the mapping of the JH1 and JH9 reads onto respectively the JH1 and JH9 contig sequences. We applied clustalw (Section 6) to the window to produce a full-multi-alignment of the JH1 and JH9 contig sequences, N315 chromosomal sequence, and JH1 and JH9 reads. We then moved onto the next window of W columns in the MAC and applied clustalw again and so on.

The final MACR had a length of about 3,000,000 columns. In the MACR, an indel in a read was assigned the Phred quality value of the previous base (Section 7.1). To identify alignment errors in the MACR, we for each of JH1 and JH9 searched for and manually examined columns in which the symbol in the JH contig sequence disagreed with a symbol in a JH read that had a Phred quality value ≥ 30. Through a process of trial and error, we were able to find optimal values for the parameters of clustalw that produced alignment errors at a rate of only about 1:200,000 columns, with no error spanning more than several columns.

To achieve this alignment error rate, the program clustalw was used as follows. It was applied to successive windows of the MAC, each containing 500 columns. When applied to a given window, it was invoked as follows:

```
clustalw -type=DNA -pwdnamatrix=bestfit_dna_matrix_less_priority_to_degenerate
        -dnamatrix=bestfit_dna_matrix_less_priority_to_degenerate -profile1=all.txt
```

-profile2=JH.txt -sequences -outfile=all.txt

The file bestfit_dna_matrix_less_priority_to_degenerate contains the following DNA scoring matrix:

	A	C	G	T	N
A	10	-9	-9	-9	5
C	-9	10	-9	-9	5
G	-9	-9	10	-9	5
T	-9	-9	-9	10	5
N	5	5	5	5	5

The scoring matrix places less emphasis on matches to the strings of N's representing contig gaps in the MAC. Initially, the file all.txt contains only the N315 chromosomal sequence, and the file JH.txt contains the JH1 and JH9 contig sequences and reads.

11. Identification and subsequent experimental verification of the mutations between the JH1, JH9, and N315 chromosomes.

11.1. Bayesian probabilistic model for identifying mutations between the JH1, JH9, and N315 chromosomes

The Bayesian probabilistic model (BPM) for identifying real ND's considered in each column in the MACR in Section 10 the symbols in the trimmed JH1 and JH9 reads and their Phred quality values. A symbol in a trimmed JH1 or JH9 read could be an A, C, G, T, N, or –, where N represents any of the four bases and – is an indel that was inserted during the multi-alignment. In each of JH1 and JH9, the four-fold degenerate symbol N occurred in the trimmed reads with a frequency of about 1:5000 nucleotides, and there was no instance in any trimmed read of a two-fold degenerate symbol representing any of two of the bases (e.g. W = A or T) or a three-fold degenerate symbol representing any of three of the bases (e.g. H = A, C, or T). Since a N serves as a placeholder and is otherwise uninformative, instances of N's were ignored. If for example a column contained six reads for JH1 and two of these reads contained a N, then the coverage in the column for JH1 would be reported as 4X, and only the four reads containing a non-N would be considered. As described in Section 10, a – in a read was assigned the quality value of the previous base in the read.

11.1.1. Use of the unambiguous mapping of the trimmed reads to contigs

The MACR in Section 10 was generated using the unambiguous mapping of the trimmed reads to contigs in Section 9.4, which excluded reads that matched more than one region. The omission of these reads was actually considered to be advantageous for the following reasons. While the exclusion of the reads ultimately reduced the coverage, the affect was confined to only repetitive sequence. When the reads were included, the coverage in JH1 and JH9 increased in only 2.2 and 1.4% respectively of the columns in the MACR, all of which were found to fall in highly repetitive regions. In these columns, the mean coverage rose from 6X to 16X in JH1 and from 8X to 17X in JH9. However, cursory examination of these columns suggested that the extra coverage was unreliable due to unresolved repeats. In some of the columns, the symbols in the reads for JH1 and/or JH9 were found be a mixture (e.g. Seven reads for JH1 contained an A, and

another six reads for JH1 contained a C.). Such mixtures are a sign of unresolved repeats. For instance, a segment of a chromosome may have an A at some position along its length whereas a near-identical copy of the segment elsewhere in the chromosome may have a C at the position. Reads from the first copy mapping to the second copy might produce a mixture of A's and C's in the column in the MACR corresponding to the stated position. Depending on the relative coverage of the two copies, a probabilistic model may therefore incorrectly call an A instead of C in the second copy at the stated position. Due to incorrectly mapped repeats, the Celera assembler is known to make errors more frequently in repetitive regions.

11.1.2. Read error rates

For S = JH1 or JH9, $X \in \{\text{A,C,G,T,}-\}$, and $Y \in \{\text{A,C,G,T,}-\}$, we define $\mathrm{E}_S(Y \mid X;Q)$ as the probability in the isolate S that the base calling program will call a symbol Y in a read given that the correct symbol is X and the call will be assigned a Phred quality value Q. Note that each of X and Y can take on five values: one of the four bases or an indel. The case $X \neq Y$ represents a read error.

As described in Section 7.1, the Phred quality value Q assigned to a symbol Y in a read is an integer from 0 to 60 defined such that $10^{-Q/10}$ is probability that Y is an incorrect call. Extensive work has shown that the Phred quality value tracks the read error reasonably well (67). Let $\mathrm{P}_S(X \mid Y;Q)$ denote the probability in the isolate S that the correct symbol is X given that the symbol called in a read was Y and was assigned a Phred quality value Q. In terms of $\mathrm{P}_S(X \mid Y;Q)$, the definition of Q is equivalent to

$$\sum_{\substack{X \\ X \neq Y}} \mathrm{P}_S(X \mid Y;Q) \equiv 10^{-Q/10}. \tag{10}$$

Using Bayes' theorem, $\mathrm{E}_S(Y \mid X;Q)$ can be related to $\mathrm{P}_S(X \mid Y;Q)$:

$$\frac{\mathrm{E}_S(Y \mid X;Q)\mathrm{P}(X)}{\sum_X \mathrm{E}_S(Y \mid X;Q)\mathrm{P}(X)} = \mathrm{P}_S(X \mid Y;Q) \tag{11}$$

where $\mathrm{P}(X)$ is the prior probability over X. In the absence of prior information, we took an uninformative prior $\mathrm{P}(X) = 1/5$, so Eq. 11 reduced to

$$\frac{\mathrm{E}_S(Y \mid X;Q)}{\sum_X \mathrm{E}_S(Y \mid X;Q)} = \mathrm{P}_S(X \mid Y;Q). \tag{12}$$

Summing over $X \neq Y$, it can be seen using Eq. 10 that the definition of Q is tantamount to the statement

$$\frac{\sum_{\substack{X \\ X \neq Y}} \mathrm{E}_S(Y \mid X;Q)}{\sum_X \mathrm{E}_S(Y \mid X;Q)} \equiv 10^{-Q/10}. \tag{13}$$

Initially, we computed $\mathrm{E}_S(Y \mid X;Q)$ not using the definition of Q but directly from the MACR by enumerating read errors. We considered only columns in the MACR in which there was high coverage for S and all but a small percentage of the symbols in the reads for S agreed. The anomalous symbols were assumed to be the result of read errors. Unfortunately, the computed $\mathrm{E}_S(Y \mid X;Q)$ were unreliable for $Q \geq$ about 30 due to insufficient counts for $X \neq Y$. Several techniques were tried to improve the estimates of the $\mathrm{E}_S(Y \mid X;Q)$ at higher Q: columns with lower coverage for S were also considered to increase counts; curves were fit to the $\mathrm{E}_S(Y \mid X;Q)$ at lower Q, and the curves were then extrapolated to estimate the $\mathrm{E}_S(Y \mid X;Q)$ at higher Q; counts were binned to compute the $\mathrm{E}_S(Y \mid X;Q)$ at higher Q; etc. However, each approach that was tried seemed to introduce sizable errors.

Although the $\mathrm{E}_S(Y \mid X;Q)$ computed directly from the MACR were unreliable, several trends did emerge: $\mathrm{E}_S(Y \mid X;Q)$ seemed to be independent of S; $\mathrm{E}_S(Y \mid X;Q)$ for $X \neq Y$ appeared to be roughly independent of X and Y over most of the range of Q; and finally, $\mathrm{E}_S(Y \mid X;Q)$ was observed to a good approximation to satisfy Eq. 13. In accordance with these observations, we made the approximation

$$\mathrm{E}_S(Y \mid X;Q) = \begin{cases} 1-\mathrm{E}(Q), & Y = X \\ \mathrm{E}(Q) \times 1/4, & Y \neq X \end{cases} \tag{14}$$

where the probability $\mathrm{E}(Q)$ that X is called incorrectly in a read as $Y \neq X$ depends only on Q. We then substituted Eq. 14 into Eq. 13 to yield

$$\mathrm{E}(Q) = 10^{-Q/10}. \tag{15}$$

Thus, Eq. 14 could be written as

$$\mathrm{E}_S(Y \mid X;Q) = \begin{cases} 1-10^{-Q/10}, & Y = X \\ 10^{-Q/10} \times 1/4, & Y \neq X \end{cases}. \tag{16}$$

We computed $\mathrm{E}_S(Y \mid X;Q)$ using Eq. 16.

11.1.3. The i^{th} column in the MACR

We considered in the MACR the i^{th} column. Like every column in the MACR, the i^{th} column contains a single symbol from N315 and symbols from the JH1 and JH9 reads. Let $Z_{\text{N315}} \in \{\text{A,C,G,T,}-\}$ denote the symbol from N315, which we assumed to be correct. Using the reads, we wished to evaluate the probability that the correct symbols in JH1 and JH9 are $X_{\text{JH1}} \in \{\text{A,C,G,T,}-\}$ and $X_{\text{JH9}} \in \{\text{A,C,G,T,}-\}$ respectively.

11.1.4. Priors

We worked with two priors over X_{JH1} and X_{JH9}:

$$\mathrm{P}(X_{\mathrm{JH1}}, X_{\mathrm{JH9}}; \alpha) = \begin{cases} \dfrac{1}{25}, & \alpha = 1 \text{ (uniformative)} \\ \dfrac{1}{5}\mathrm{M}_{\mathrm{JH1,JH9}}\left(X_{\mathrm{JH9}} \mid X_{\mathrm{JH1}}\right), & \alpha = 2 \text{ (phylogenetic)} \end{cases} . \quad (17)$$

In the uninformative prior $\alpha = 1$, each of the 25 combinations of values of X_{JH1} and X_{JH9} are equally likely. The phylogenetic prior $\alpha = 2$ reflects the phylogeny between JH1 and JH9, as defined by the point mutation rate $\mathrm{M}_{\mathrm{JH1,JH9}}(Y \mid X)$.

We defined $\mathrm{M}_{\mathrm{JH1,JH9}}(Y \mid X)$ to be the probability that a symbol X in the isolate JH1 was changed to the symbol Y in the isolate JH9 by a point mutation. In the special case $X = Y$, $\mathrm{M}_{\mathrm{JH1,JH9}}(Y = X \mid X)$ is defined as the probability of no point mutation occurring. We computed $\mathrm{M}_{\mathrm{JH1,JH9}}(Y \mid X)$ as follows:

$$\mathrm{M}_{\mathrm{JH1,JH9}}(Y \mid X) = \begin{cases} 1 - m_{\mathrm{JH1,JH9}}, & Y = X \\ m_{\mathrm{JH1,JH9}} \dfrac{1}{4}, & Y \neq X \end{cases} \quad (18)$$

where $m_{\mathrm{JH1,JH9}}$ is some estimate that needs to be determined of the overall point mutation rate between JH1 and JH9 that includes insertions, deletions, and substitutions. It is assumed that the point mutation rate $\mathrm{M}_{\mathrm{JH1,JH9}}(Y \mid X)$ is independent of X and Y when $X \neq Y$. A more refined model is not possible for two reasons. Firstly, the dependence of $\mathrm{M}_{\mathrm{JH1,JH9}}(Y \mid X)$ on X and Y when $X \neq Y$ may vary widely from locus to locus. The strong selective pressures due to antibiotic chemotherapy can select for a particular type of point mutation in one locus and a different type of point mutation in another locus. Secondly, the dependence cannot be reliably computed since the mutations between JH1 and JH9 are so rare.

In both priors $\alpha = 1$ or 2, we did not consider the phylogeny between N315 and the JH lineage because N315 and the JH lineage differ by large regions of non-homology (including one replacement > 40,000 nucleotides) and large inserts and deletions (including three > 40,000 nucleotides). Thus, the phylogeny between N315 and the JH lineage cannot be characterized by a simple prior that assumes a single mutation rate and ignores spatial correlations by assuming all differences are due to independent point mutations. Because the mutations between JH1 and JH9 are so rare, we could not a priori reliably estimate the mutation rate between JH1 and JH9, so we used the uninformative prior $\alpha = 1$ to predict real ND's. Once we had a reliable estimate of the mutation rate between JH1 and JH9 in the regions of high coverage and good quality, we used the phylogenetic prior $\alpha = 2$ to estimate the number of mutations in the regions of low coverage and poor quality.

11.1.5. Conditional probabilities

We computed the conditional probability $\mathrm{P}(i \mid X_{\mathrm{JH1}}, X_{\mathrm{JH9}})$ of observing in the i^{th} column the reads given that the correct symbols in JH1 and JH9 are X_{JH1} and X_{JH9} respectively:

$$\mathrm{P}(i \mid X_{\mathrm{JH1}}, X_{\mathrm{JH9}}) = \left[\prod_{\substack{Y,Q \\ \text{over JH1 reads}}} \mathrm{E}_{\mathrm{JH1}}(Y \mid X_{\mathrm{JH1}}; Q) \right] \left[\prod_{\substack{Y,Q \\ \text{over JH9 reads}}} \mathrm{E}_{\mathrm{JH9}}(Y \mid X_{\mathrm{JH9}}, Q) \right]. \tag{19}$$

We used the read error probabilities in Eq. 16.

11.1.6. Posterior probabilties

Using Bayes' theorem, we computed the posterior probability $\mathrm{P}(X_{\mathrm{JH1}}, X_{\mathrm{JH9}} \mid i; \alpha)$ that the correct symbols in the i^{th} column in JH1 and JH9 are X_{JH1} and X_{JH9} respectively given the observed reads:

$$\mathrm{P}(X_{\mathrm{JH1}}, X_{\mathrm{JH9}} \mid i; \alpha) = \frac{\mathrm{P}(i \mid X_{\mathrm{JH1}}, X_{\mathrm{JH9}}) \mathrm{P}(X_{\mathrm{JH1}}, X_{\mathrm{JH9}}; \alpha)}{\sum_{\substack{U \in \{\mathrm{A,C,G,T,-}\} \\ V \in \{\mathrm{A,C,G,T,-}\}}} \mathrm{P}(i \mid U, V) \mathrm{P}(U, V; \alpha)}. \tag{20}$$

We used the prior α in Eq. 17 and the conditional probabilities in Eq. 19. We then computed the posterior probability $\mathrm{P}(\mathrm{JH1} \neq \mathrm{N315} \mid i; \alpha)$ of a real ND in the i^{th} column between JH1 and N315:

$$\mathrm{P}(\mathrm{JH1} \neq \mathrm{N315} \mid i; \alpha) = \sum_{\substack{X_{\mathrm{JH1}} \\ X_{\mathrm{JH1}} \neq Z_{\mathrm{N315}}}} \sum_{X_{\mathrm{JH9}}} \mathrm{P}(X_{\mathrm{JH1}}, X_{\mathrm{JH9}} \mid i; \alpha). \tag{21}$$

We also computed the posterior probability $\mathrm{P}(\mathrm{JH1} \neq \mathrm{JH9} \mid i; \alpha)$ of a real ND in the i^{th} column between JH1 and JH9:

$$\mathrm{P}(\mathrm{JH1} \neq \mathrm{JH9} \mid i; \alpha) = \sum_{X_{\mathrm{JH1}}} \sum_{\substack{X_{\mathrm{JH9}} \\ X_{\mathrm{JH9}} \neq X_{\mathrm{JH1}}}} \mathrm{P}(X_{\mathrm{JH1}}, X_{\mathrm{JH9}} \mid i; \alpha). \tag{22}$$

11.1.7. Identification of real ND's and informative columns

For (S_1, S_2) = (N315, JH1) or (JH1, JH9), we compared the two strains S_1 and S_2. In the absence of prior information, we used the uninformative prior $\alpha = 1$ (see Eq. 17). The posterior probability $\mathrm{P}(S_1 \neq S_2 \mid i; \alpha = 1)$ of a real ND in the i^{th} column between S_1 and S_2 (see Eq.'s 21 and 22) must satisfy one and only one of the three following conditions:

$$\mathrm{P}(S_1 \neq S_2 \mid i; \alpha = 1) > T \qquad \text{(case 1, informative, real ND)} \tag{23}$$

or

$$1 - T \leq \mathrm{P}(S_1 \neq S_2 \mid i; \alpha = 1) \leq T \qquad \text{(case 2, uninformative, no call)} \tag{24}$$

or

$$\mathrm{P}(S_1 \neq S_2 \mid i; \alpha = 1) < 1 - T \qquad \text{(case 3, informative, no real ND)} \tag{25}$$

where

$$T = 1 - 1/\left(3 \times 10^6\right). \tag{26}$$

In case 1, we labeled the i^{th} column informative and predicted a real ND; in case 2, we labeled the column uninformative and made no call either way about a real ND; and in case 3, we labeled the column informative and ruled out a real ND. The threshold $T = 1 - 1/\left(3 \times 10^6\right)$ was

selected so that the expected error (both false positive and false negative) for calling a real ND is less than one for the entire MACR.

When a real ND was predicted in an informative column, the region containing the column in the MACR was always manually examined. The 10-20 columns with alignment errors in the MACR were checked (see Section 10). Anomalous predictions due to alignment errors were identified and not reported. It was also ensured that no real ND went unreported due to an alignment error.

11.2. Reported mutations between the JH1 and N315 chromosomes

In the comparison of JH1 and N315, 97% of the columns in the MACR were informative, and we reported only the predicted real ND's in informative columns. Runs of real ND's occurring in regions of non-homology or arising due to large inserts and deletions were identified as such. The results are summarized in Table S.3 and Figures S.1(a) and S.2.

11.3. Reported mutations between JH1 and JH9 chromosomes

In the JH1 and JH9 comparison, 94% of the columns in the MACR were informative. PCR sequencing was done to check all the predicted real ND's in the informative columns. As already stated, all of the predicted real ND's were confirmed except one. In the sole exceptional case, the PCR sequencing method failed, and the result was inconclusive. Also, PCR sequencing was done to check the 10 uninformative columns with the highest $P(\mathrm{JH1} \neq \mathrm{JH9} \mid i; \alpha = 1)$, that is the 10 uninformative columns most likely to contain a real ND. This amounted to checking all uninformative columns with a $P(\mathrm{JH1} \neq \mathrm{JH9} \mid i; \alpha = 1) \geq 1 - 10^{-4}$. In eight of these columns, a real ND was ruled out. For one column, the PCR sequencing method failed, and the result was inconclusive. In the remaining column, a real ND was found. See Table M.6 for a summary of the $P(\mathrm{JH1} \neq \mathrm{JH9} \mid i; \alpha = 1)$ scores.

Table M.6. An overview of the $P(\mathrm{JH1} \neq \mathrm{JH9} \mid i; \alpha = 1)$ **scores.** Here, $P(\mathrm{JH1} \neq \mathrm{JH9} \mid i; \alpha = 1)$ is denoted simply by P.

34 informative columns with a predicted real ND [P > 1 – 1/(3×10^6)]
• P ranges from $1-10^{-80}$ to $1-10^{-11}$. • PCR sequencing was done to check all columns. All were confirmed to contain a real ND except one. In the sole exceptional case, the PCR sequencing method failed, and the result was inconclusive.
Uniformative columns [1/(3×10^6) ≤ P ≤ 1 – 1/(3×10^6)]
10 uniformative columns for which P ≥ 1–10^{-4}
• PCR sequencing was done to check all columns. In eight of these columns, a real ND was ruled out. For one column, the PCR sequencing method failed, and the result was inconclusive. In the remaining column, a real ND was found.
Remaining 176,994 uniformative columns for which P < 1–10^{-4}
• Shown in bold are the three columns with the highest P's (i.e. the 3 columns most likely to contain a real ND). Also shown are adjacent columns. Differences in the two columns with the highest P's occurred in the same run of five T's at the end of a read. Poly-A and -T sequence at the end of a read is known to be unreliable.

two columns with highest P's

JH1 6X coverage	A	G	A	G	A	T	**C**	**A**	T	T	T	A	T	T	G	–
	A	G	G	G	A	T	**C**	**A**	T	T	T	A	T	T	G	–
	A	G	G	G	A	T	**C**	**A**	T	T	T	A	T	T	G	–
	A	G	G	G	A	T	**C**	**A**	T	T	T	A	T	T	G	G
	A	G	G	G	A	T	**C**	**A**	T	T	T	A	T	T	G	–
	A	G	G	G	A	T	**C**	**A**	T	T	T	A	T	T	G	–
JH9 1X coverage	A	G	G	G	A	T	**T**	**T**	T	T	read ended now 0X coverage					
							P = 1–10$^{-3.9}$	**P = 1–10$^{-3.9}$**								

third column

JH1 2X coverage	T	C	A	A	A	**A**	G	A	T	T	G
	T	C	A	A	A	**A**	G	A	T	T	G
JH9 1X coverage	T	C	A	A	A	**–**	G	A	T	T	T
						P = 1–10$^{-3.8}$					

- 90,115 columns have 0X coverage, 64,169 columns have 1X coverage, and the remaining 176,994 –90,115–64,169 = 22710 have almost exclusively poor quality 2X coverage in JH1 or JH9 or both. Thus, there is insufficient read information to discriminate between read ND's and read errors.

- Of course, the 90,115 columns with 0X coverage in JH1 or JH9 or both all have a P = 4/5, corresponding to the a priori probability of a real ND.

- Only 458 columns have a P > 4/5. The vast majority of the differences in these columns are expected to be read errors. Many of the differences occur in 1X coverage in poly-A or –T sequence at the end of a read, which is known to be unreliable. A read error rate between 10^{-2} and 10^{-3} could easily generate in 1X coverage on the order of 500 read errors in regions totaling 60,000 bases.

Informative columns for which a real ND was ruled out [P < 1/(3×10^6)]

- P ranges from $10^{-6.5}$ to 10^{-110}

- Shown in bold is the column with the highest P (i.e. the column most likely to contain a real ND). Also shown are adjacent columns.

JH1 4X coverage	T	A	A	A	T	**G**	A	C	A	C	A
	T	A	A	A	T	**G**	A	C	A	C	A
	T	A	A	A	T	**G**	A	C	A	C	A
	T	A	A	A	T	**G**	A	C	A	C	A
JH9 2X coverage	T	A	A	A	T	**G**	A	C	A	C	A
	T	A	A	A	T	**G**	A	C	A	C	A
						P = 10$^{-6.5}$					

11.4. Estimation of number of unreported mutations between JH1 and JH9 chromosomes

In the JH1 and JH9 comparison, only point mutations were found in the 94% of the informative columns in the MACR. Moreover, the point mutations were observed to occur at a rate of 35 in 94% of the columns (1 per 80,000 bp). To estimate the number $N_{\text{unreported}}$ of unreported mutations between the JH1 and JH9 chromosomes, we assumed that only point mutations occurred in the remaining 6% of uninformative columns, also at a rate of 1:80,000 bp. We estimated $N_{\text{unreported}}$ as follows:

$$N_{\text{unreported}} = \frac{35}{0.94} \times 0.06 = 2.2 \,. \tag{27}$$

However, this estimation assumed that we have no information in the uninformative columns as to whether there is or is not a real ND. This is true only in the regions of 0X coverage. The regions of 1-2X coverage offer some discriminatory power. To produce a more rigorous estimate of $N_{\text{unreported}}$, we chose $m_{\text{JH1,JH9}} = 1/80{,}000$ in Eq. 18, used the phylogenetic prior $\alpha = 2$ in Eq. 17, and computed $N_{\text{unreported}}$ as follows:

$$N_{\text{unreported}} = \sum_{\substack{i \\ \text{uniformative columns for which} \\ \text{P(JH1}\neq\text{JH9}|i;\alpha=1) < 1-10^{-4}}} \text{P(JH1} \neq \text{JH9} \mid i; \alpha = 2) \qquad = 1.5 \tag{28}$$

where the sum is over the uninformative columns not checked by PCR sequencing for which $\text{P(JH1} \neq \text{JH9} \mid i; \alpha = 1) < 1 - 10^{-4}$.

References Citied in Appendix

1. Berger-Bachi, B. & Rohrer, S. (2002) *Arch Microbiol* **178,** 165-71.
2. Zhang, H. Z., Hackbarth, C. J., Chansky, K. M. & Chambers, H. F. (2001) *Science* **291,** 1962-5.
3. Hackbarth, C. J., Miick, C. & Chambers, H. F. (1994) *Antimicrob Agents Chemother* **38,** 2568-71.
4. Yin, S., Daum, R. S. & Boyle-Vavra, S. (2006) *Antimicrob Agents Chemother* **50,** 336-43.
5. Kuroda, M., Kuroda, H., Oshima, T., Takeuchi, F., Mori, H. & Hiramatsu, K. (2003) *Mol Microbiol* **49,** 807-21.
6. Kuroda, M., Kuwahara-Arai, K. & Hiramatsu, K. (2000) *Biochem Biophys Res Commun* **269,** 485-90.
7. Gardete, S., Wu, S. W., Gill, S. & Tomasz, A. (2006) *Antimicrob Agents Chemother* **50,** 3424-34.
8. O'Neill, A. J., Huovinen, T., Fishwick, C. W. & Chopra, I. (2006) *Antimicrob Agents Chemother* **50,** 298-309.
9. Wichelhaus, T. A., Boddinghaus, B., Besier, S., Schafer, V., Brade, V. & Ludwig, A. (2002) *Antimicrob Agents Chemother* **46,** 3381-5.
10. Campbell, E. A., Korzheva, N., Mustaev, A., Murakami, K., Nair, S., Goldfarb, A. & Darst, S. A. (2001) *Cell* **104,** 901-12.
11. Friedman, L., Alder, J. D. & Silverman, J. A. (2006) *Antimicrob Agents Chemother* **50,** 2137-45.
12. Maughan, H., Galeano, B. & Nicholson, W. L. (2004) *J Bacteriol* **186,** 2481-6.
13. Jin, D. J., Cashel, M., Friedman, D. I., Nakamura, Y., Walter, W. A. & Gross, C. A. (1988) *J Mol Biol* **204,** 247-61.
14. Jin, D. J., Walter, W. A. & Gross, C. A. (1988) *J Mol Biol* **202,** 245-53.
15. Jin, D. J. & Gross, C. A. (1991) *J Biol Chem* **266,** 14478-85.
16. Nickels, B. E. & Hochschild, A. (2004) *Cell* **118,** 281-4.
17. Lyon, G. J. & Novick, R. P. (2004) *Peptides* **25,** 1389-403.
18. Dunman, P. M., Murphy, E., Haney, S., Palacios, D., Tucker-Kellogg, G., Wu, S., Brown, E. L., Zagursky, R. J., Shlaes, D. & Projan, S. J. (2001) *J Bacteriol* **183,** 7341-53.
19. Korem, M., Gov, Y., Kiran, M. D. & Balaban, N. (2005) *Infect Immun* **73,** 6220-8.
20. Liang, X., Zheng, L., Landwehr, C., Lunsford, D., Holmes, D. & Ji, Y. (2005) *J Bacteriol* **187,** 5486-92.
21. Novick, R. P. (2003) *Mol Microbiol* **48,** 1429-49.
22. Sakoulas, G., Moellering, R. C., Jr. & Eliopoulos, G. M. (2006) *Clin Infect Dis* **42 Suppl 1,** S40-50.
23. Sakoulas, G., Eliopoulos, G. M., Fowler, V. G., Jr., Moellering, R. C., Jr., Novick, R. P., Lucindo, N., Yeaman, M. R. & Bayer, A. S. (2005) *Antimicrob Agents Chemother* **49,** 2687-92.
24. Sakoulas, G., Eliopoulos, G. M., Moellering, R. C., Jr., Wennersten, C., Venkataraman, L., Novick, R. P. & Gold, H. S. (2002) *Antimicrob Agents Chemother* **46,** 1492-502.
25. Traber, K. & Novick, R. (2006) *Mol Microbiol* **59,** 1519-30.
26. Dubrac, S. & Msadek, T. (2004) *J Bacteriol* **186,** 1175-81.
27. Sieradzki, K. & Tomasz, A. (2006) *Antimicrob Agents Chemother* **50,** 527-33.
28. Howell, A., Dubrac, S., Andersen, K. K., Noone, D., Fert, J., Msadek, T. & Devine, K. (2003) *Mol Microbiol* **49,** 1639-55.
29. Martin, P. K., Li, T., Sun, D., Biek, D. P. & Schmid, M. B. (1999) *J Bacteriol* **181,** 3666-73.
30. Ito, M., Guffanti, A. A., Wang, W. & Krulwich, T. A. (2000) *J Bacteriol* **182,** 5663-70.
31. Mazmanian, S. K., Skaar, E. P., Gaspar, A. H., Humayun, M., Gornicki, P., Jelenska, J., Joachmiak, A., Missiakas, D. M. & Schneewind, O. (2003) *Science* **299,** 906-9.
32. Mack, J., Vermeiren, C., Heinrichs, D. E. & Stillman, M. J. (2004) *Biochem Biophys Res Commun* **320,** 781-8.
33. Mazmanian, S. K., Ton-That, H., Su, K. & Schneewind, O. (2002) *Proc Natl Acad Sci U S A* **99,** 2293-8.
34. Vitikainen, M., Lappalainen, I., Seppala, R., Antelmann, H., Boer, H., Taira, S., Savilahti, H., Hecker, M., Vihinen, M., Sarvas, M. & Kontinen, V. P. (2004) *J Biol Chem* **279,** 19302-14.
35. Tossavainen, H., Permi, P., Purhonen, S. L., Sarvas, M., Kilpelainen, I. & Seppala, R. (2006) *FEBS Lett* **580,** 1822-6.

36. Vitikainen, M., Pummi, T., Airaksinen, U., Wahlstrom, E., Wu, H., Sarvas, M. & Kontinen, V. P. (2001) *J Bacteriol* **183,** 1881-90.
37. Hyyrylainen, H. L., Vitikainen, M., Thwaite, J., Wu, H., Sarvas, M., Harwood, C. R., Kontinen, V. P. & Stephenson, K. (2000) *J Biol Chem* **275,** 26696-703.
38. Hyyrylainen, H. L., Sarvas, M. & Kontinen, V. P. (2005) *Appl Microbiol Biotechnol* **67,** 389-96.
39. Sieradzki, K. & Tomasz, A. (2003) *J Bacteriol* **185,** 7103-10.
40. Wei, Y., Guffanti, A. A., Ito, M. & Krulwich, T. A. (2000) *J Biol Chem* **275,** 30287-92.
41. McAleese, F., Wu, S. W., Sieradzki, K., Dunman, P., Murphy, E., Projan, S. & Tomasz, A. (2006) *J Bacteriol* **188,** 1120-33.
42. Komatsuzawa, H., Fujiwara, T., Nishi, H., Yamada, S., Ohara, M., McCallum, N., Berger-Bachi, B. & Sugai, M. (2004) *Mol Microbiol* **53,** 1221-31.
43. Dugourd, D., Martin, C., Rioux, C. R., Jacques, M. & Harel, J. (1999) *J Bacteriol* **181,** 6948-57.
44. Sieradzki, K., Leski, T., Dick, J., Borio, L. & Tomasz, A. (2003) *J Clin Microbiol* **41,** 1687-93.
45. Kuroda, M., Ohta, T., Uchiyama, I., Baba, T., Yuzawa, H., Kobayashi, I., Cui, L., Oguchi, A., Aoki, K., Nagai, Y., et al. (2001) *Lancet* **357,** 1225-40.
46. Sieradzki K, R. R., Haber SW, Tomasz A. (1999) *N Engl J Med*. **340,** 517-23.
47. (1997) *MMWR Morb Mortal Wkly Rep* **46,** 765-756.
48. Sieradzki, K. & Tomasz, A. (1999) *J Bacteriol* **181,** 7566-70.
49. Enright, M. C., Day, N. P., Davies, C. E., Peacock, S. J. & Spratt, B. G. (2000) *J Clin Microbiol* **38,** 1008-15.
50. Shopsin, B., Gomez, M., Montgomery, S. O., Smith, D. H., Waddington, M., Dodge, D. E., Bost, D. A., Riehman, M., Naidich, S. & Kreiswirth, B. N. (1999) *J Clin Microbiol* **37,** 3556-63.
51. http://www.ncbi.nlm.nih.gov/genomes/static/eub_g.html.
52. Gill, S. R., Fouts, D. E., Archer, G. L., Mongodin, E. F., Deboy, R. T., Ravel, J., Paulsen, I. T., Kolonay, J. F., Brinkac, L., Beanan, M., Dodson, R. J., Daugherty, S. C., Madupu, R., Angiuoli, S. V., Durkin, A. S., Haft, D. H., Vamathevan, J., Khouri, H., Utterback, T., Lee, C., Dimitrov, G., Jiang, L., Qin, H., Weidman, J., Tran, K., Kang, K., Hance, I. R., Nelson, K. E. & Fraser, C. M. (2005) *J Bacteriol* **187,** 2426-38.
53. Drake, J. W., Charlesworth, B., Charlesworth, D. & Crow, J. F. (1998) *Genetics* **148,** 1667-86.
54. Ito, T., Katayama, Y. & Hiramatsu, K. (1999) *Antimicrob Agents Chemother* **43,** 1449-58.
55. Kwan, T., Liu, J., DuBow, M., Gros, P. & Pelletier, J. (2005) *Proc Natl Acad Sci U S A* **102,** 5174-9.
56. Iandolo, J. J., Worrell, V., Groicher, K. H., Qian, Y., Tian, R., Kenton, S., Dorman, A., Ji, H., Lin, S., Loh, P., Qi, S., Zhu, H. & Roe, B. A. (2002) *Gene* **289,** 109-18.
57. Diep, B. A., Gill, S. R., Chang, R. F., Phan, T. H., Chen, J. H., Davidson, M. G., Lin, F., Lin, J., Carleton, H. A., Mongodin, E. F., Sensabaugh, G. F. & Perdreau-Remington, F. (2006) *Lancet* **367,** 731-9.
58. Murphy, E., Huwyler, L. & de Freire Bastos Mdo, C. (1985) *Embo J* **4,** 3357-65.
59. Derbise, A., Dyke, K. G. & el Solh, N. (1994) *Plasmid* **31,** 251-64.
60. Chesneau, O., Lailler, R., Derbise, A. & El Solh, N. (1999) *FEMS Microbiol Lett* **177,** 93-100.
61. Ohta, T., Hirakawa, H., Morikawa, K., Maruyama, A., Inose, Y., Yamashita, A., Oshima, K., Kuroda, M., Hattori, M., Hiramatsu, K., Kuhara, S. & Hayashi, H. (2004) *DNA Res* **11,** 51-6.
62. Venter, J. C., Smith, H. O. & Hood, L. (1996) *Nature* **381,** 364-6.
63. Venter, J. C., Adams, M. D., Sutton, G. G., Kerlavage, A. R., Smith, H. O. & Hunkapiller, M. (1998) *Science* **280,** 1540-2.
64. Myers, E. W., Sutton, G. G., Delcher, A. L., Dew, I. M., Fasulo, D. P., Flanigan, M. J., Kravitz, S. A., Mobarry, C. M., Reinert, K. H., Remington, K. A., Anson, E. L., Bolanos, R. A., Chou, H. H., Jordan, C. M., Halpern, A. L., Lonardi, S., Beasley, E. M., Brandon, R. C., Chen, L., Dunn, P. J., Lai, Z., Liang, Y., Nusskern, D. R., Zhan, M., Zhang, Q., Zheng, X., Rubin, G. M., Adams, M. D. & Venter, J. C. (2000) *Science* **287,** 2196-204.
65. Hohl, M., Kurtz, S. & Ohlebusch, E. (2002) *Bioinformatics* **18 Suppl 1,** S312-20.

66. Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994) *Nucleic Acids Res* **22,** 4673-80.
67. Ewing, B. & Green, P. (1998) *Genome Res* **8,** 186-94.
68. Morgenstern, B. (2004) *Nucleic Acids Res* **32,** W33-6.
69. http://www.jgi.doe.gov/.
70. Rasmussen, K. R., Stoye, J. & Myers, E. W. (2006) *J Comput Biol* **13,** 296-308.
71. Szurmant, H., Nelson, K., Kim, E. J., Perego, M. & Hoch, J. A. (2005) *J Bacteriol* **187,** 5419-26.
72. McGinnis, S. & Madden, T. L. (2004) *Nucleic Acids Res* **32,** W20-5.
73. Detter, J. C., Jett, J. M., Lucas, S. M., Dalin, E., Arellano, A. R., Wang, M., Nelson, J. R., Chapman, J., Lou, Y., Rokhsar, D., Hawkins, T. L. & Richardson, P. M. (2002) *Genomics* **80,** 691-8.