Molecular investigations on a chimeric strain of *Staphylococcus aureus* sequence type 80

Darius Gawlik [1,2] (d.gawlik@ptc-phage.com)

Antje Ruppelt-Lorz [3] (antje.ruppelt@gmx.de)

Elke Müller [4,6] (elke.mueller@leibniz-ipht.de)

Annett Reißig [4,6] (annett.reissig@leibniz-ipht.de)

Helmut Hotzel [7] (helmut.hotzel@fli.de)

Sascha D. Braun [4,6] (sascha.braun@leibniz-ipht.de)

Bo Söderquist [8] (bo.soderquist@regionorebrolan.se)

Albrecht Ziegler-Cordts [9] (ziegler.albrecht@gmail.com)

Ralf Ehricht [4,5,6] (ralf.ehricht@leibniz-ipht.de)

Stefan Monecke[3,4,6] (stefan.monecke@leibniz-ipht.de) *

*Corresponding author

Author details

[1] Institute of Infectious Diseases and Infection Control, University Hospital, Jena, Germany

[2] PTC - Phage Technology Center GmbH, Bönen

[3] Institute for Medical Microbiology and Hygiene, Technical University of Dresden, Dresden, Germany

[4] Leibniz Institute of Photonic Technology (IPHT), Jena, Germany

[5] Institute of Physical Chemistry, Jena University, Jena, Germany

[6] InfectoGnostics Research Campus, Jena, Germany

[7] Friedrich-Loeffler-Institut, Institute of Bacterial Infections and Zoonoses, Jena, Germany

[8] School of Medical Sciences, Department of Laboratory Medicine, Clinical Microbiology, Faculty of Medicine and Health, Örebro University, SE-701 82 Örebro, Sweden

[9] T-Systems Multimedia Solutions GmbH, Dresden, Germany

Keywords : *Staphylococcus aureus*, clonal complex 80, microarray, next generation sequencing, horizontal gene transfer, recombination

Abstract

An Eritrean patient was admitted with suspected tuberculous cervical lymphadenitis. While no mycobacteria were detected in pus from this process, culture yielded PVL-positive, methicillin-susceptible *Staphylococcus aureus*. Microarray hybridisation assigned the isolate to clonal complex (CC) 80 but revealed unusual features, including the presence of the ORF-CM14 enterotoxin homologue and of an ACME-III element as well as the absence of *etD* and *edinB*. The isolate was subjected to both, Illumina and Nanopore sequencing allowing characterisation of deviating regions within the strain´s genome. Atypical features of this strain were attributable to the presence of two genomic regions that originated from other *S. aureus* lineages and that comprised, respectively, 3% and 1.4% of the genome. One deviating region extended from *walJ* to *sirB*. It comprised ORF-CM14 and the ACME-III element. A homologous, but larger fragment was also found in an atypical *S. aureus* CC1/ST567 strain whose lineage might have served as donor of this genomic region. This region itself is a chimera comprising fragments from CC1 as well as fragments of unknown origin. The other region of another 3% of the genome comprised the region from *htsB* to *ecfA2*. It was very similar to CC1 sequences. This suggests either an incorporation of CC1 DNA into the study strain, or it might alternatively suggest a recombination event affecting "canonical" CC80. As the study strain bears witness of several recombination events, such complex and large-scale events cannot be rare and exceptional, despite a mainly clonal nature of *S. aureus*. Although the exact mechanism is not yet clear, chimerism seems to be an additional pathway in the evolution of *S. aureus*, possibly being responsible for the transmission also of virulence and resistance factors. An organism that can shuffle, swap or exchange major parts of its genome by a yet unknown mechanism would have an evolutionary advantage compared to a strictly clonal organism.

Introduction

Staphylococcus aureus (*S. aureus*) is a versatile pathogen that colonises or infects a large fraction of the world´s human population as well as several species of wild and domestic animals. Thus, it can asymptomatically colonise its carriers, or alternatively cause various infections ranging from superficial skin and soft tissue infections to serious bacteremia including infective endocarditis. Many of its virulence factors are variably present and their genes are localized on mobile genetic elements such as plasmids, phages, transposons or on pathogenicity islands. In recent decades, some strains of *S. aureus* acquired resistance to many or most antibiotics. Again, resistance genes are localized on mobile, or potentially mobile, genetic elements such as staphylococcal chromosomal cassette *mec* (SCC*mec*) cassettes. Despite a vast variety of variable, mobile elements, and despite some incremental, mutation-driven variation, the overall structure of the *S. aureus* genome is conservative with all core genomic elements being present in all strains in one uniform sequential arrangement. Multilocus sequence typing (MLST) enables the unambiguous assignment of isolates to taxonomic units, sequence types (ST) and clonal complexes (CC), based on numbered alleles of seven housekeeping genes assuming that these genes cannot be lost or truncated because of their crucial function and that the accumulation of mutations in their sequences is purely a function of time. This lead to a model of a clonal evolution of the *S. aureus* core genome that is driven by a time-dependent accumulation of single point mutations allowing classification based on a few marker genes into a limited number of clonal complexes comprising a number of sequence types that differ only in random mutations in these marker genes as well as of others.

However, some *S. aureus* strains show features that cannot be explained neither by accumulation of single point mutations nor by acquisition or loss of mobile genetic elements. For instance, there is a Russian CC8 strain in which nearly half of the genome is inverted [1].

Other strains show evidence of large-scale recombination events, with considerably large fragments of their genomes originating from other lineages and being inserted at the appropriate position of the recipient strain. Such a phenomenon was first postulated for ST239 strains, in which deviant alleles of *arcC* and *spa*, *aur, clfB* and *isaB,* of the capsule type (5 instead of 8) and the presence of *cna* indicate an integration of a CC30 DNA fragment of approximately 635,000 base pairs (or ca. 20% of the genome) into a CC8 recipient strain, with the integration site being localised around *oriC* [2]. Another strain, ST2249, harbours ST239 DNA comprising fragments of both, CC8 and CC30 that is integrated into a CC45 genome [3]. Further examples for chimeric strains are ST34 and ST42 (in which CC10 fragments are integrated into CC30 genomes) [2] or CC398 strains that harbour fragments of CC9 origin [4, 5]. Such observations indicate that large scale recombination events played a role in driving the evolution of *S. aureus* but the underlying molecular mechanisms are not yet described.

In the present paper, we describe another, new chimeric strain that comprises a backbone of CC80 genomic DNA and two separate large inserts that attracted attention because of a presence of ORF-CM14 and an absence of *edinB* and *etD*.

S. aureus CC80 is a lineage that is commonly associated with recurrent and/or severe skin and soft tissue infections (SSTI) since a majority of isolates carries the phage-borne Panton-Valentine leukocidin (PVL), a virulence factor associated with SSTI. One PVL-positive, methicillin-resistant CC80 strain, with a SCC*mec* IVc element, is sporadically found in Western Europe [6-14] while it is widespread in Mediterranean countries including Greece [15, 16], Turkey [17], Lebanon [18], Malta [19], Tunisia [20, 21] and Algeria [22-24] as well as in the Middle East/Arabian Gulf regions [25-28]. It can also commonly be found in European travellers to these regions [10, 17]. Methicillin-susceptible CC80 strains are uncommon but geographically widespread in Africa [29-31] from where this lineage originated [14].

The isolate described herein was initially subjected to microarray hybridisation, primarily for typing and detection resistance and toxin genes. The procedure revealed unusual features that could be explained by a large-scale horizontal gene transfer. This observation prompted further investigations including Illumina and Nanopore sequencing in order to map the entire genome of the isolate and a search for the donor strain of regions assumed to be introduced by the gene transfer.

Material and Methods

Clinical background and isolates

An Eritrean patient was admitted in 2015 to the Dresden University Hospital (Dresden, Saxony, Germany) with a cervical skin and soft tissue infection that originally was suspected to be suppurative tuberculous lymphadenitis. While no mycobacteria were detected (neither immediately by microscopy after staining for acid-fast bacilli nor subsequently in MGIT and Ogawa cultures), culture of pus from this process yielded a PVL-positive, methicillin-susceptible *S. aureus* (ANR570100).

A second isolate was further characterised because of certain similarities with the study isolate (see below). It was isolated in 2002. It originated from an approximately 50 years old female patient with lobar pneumonia probably secondary to an influenza B infection. She was a Swedish citizen and denied any traveling outside Sweden.

Microarray-based typing

The *S. aureus* isolates were initially characterized using different DNA microarray-based assays. Probes, primers as well as amplification and hybridization protocols have previously been described in detail [32-34].

Illumina sequencing

Genomic DNA for whole-genome sequencing was prepared from culture on Columbia blood agar incubated overnight at 37°C. DNA was prepared using the Qiagen DNA isolation kit (Qiagen, Hilden, Germany) according to manufacturer's instructions after an enzymatic lysis step with lysostaphin, lysozyme and RNAse as previously described [32-34]. Afterwards, whole-genome sequencing was carried out using the Illumina HiSeq2500 genome analyser

(Illumina HiSeq 2500 platform, Illumina, Essex, UK). The reads were assembled to contigs using SPAdes.

Sequencing of the two strains was performed at two geographically distant facilities and at different dates (Jena, Germany, and Örebro, Sweden, in spring and autumn 2018, respectively), ruling out any possibility of carry-over contaminations.

Nanopore sequencing

The Nanopore Oxford MinION platform was used for whole-genome sequencing. Briefly, no size selection was performed and the DNA library was generated using the native barcoding expansion kit EXP-NBD103 and the Nanopore sequencing kit SQK-LSK109 (Oxford Nanopore Technologies, Oxford, UK) according to manufacturer's instructions. The used flow cell FLO-MIN106 (R9-Version) was primed by the flow cell priming kit EXP-FLP001 (Oxford Nanopore). The protocol "1D Native barcoding genomic DNA" was used in version NBE_9065_v109_revB_23May2018 (Last update: 03/09/2018). The albacore basecaller (Oxford Nanopore) translated the minion raw data (FAST5) into short quality tagged sequence reads (FASTQ). After barcode trimming using Poreshop (https://github.com/rrwick/Porechop, release v0.2.4), canu (https://github.com/marbl/canu, release v1.7.1) was used to assemble the short reads. After nano-polishing (https://github.com/jts/nanopolish, release v0.11.3), the corrected sequence data were used for a direct comparison to the Illumina sequence data (see below).

Bioinformatic analysis

Iterated BLAST searches were used for analysis of individual contigs in this genome (https://blast.ncbi.nlm.nih.gov/Blast.cgi). This analysis was conducted using automated scripts for full text comparison and BLAST analysis and an in-house database of known, annotated and previously identified *S. aureus* genomes, genes and fragments to the query

sequence. This enables the determination of identity, gene content, clonal parentage and of position within the genome of each contig given the constant order of core genomic genes in *S. aureus*.

Contigs were compared to the CC80 strain 11819-97, GenBank CP003194.1/SAMN02603886. This is a PVL-positive strain with a SCC*mec* IV element and – as essentially all canonical CC80 strains – with an *etD/edinB* pathogenicity island. It can be regarded as representative for CC80. Its genome has a size of 2,846,546 nt and an average G/C content of 32.9%. They were also compared to the long-known CC8 strain COL, GenBank CP000046, and to MW2, BA000033, as reference sequence for CC1.

Finally, Nanopore and Illumina sequences were aligned manually for reasons discussed below.

Results

Comparison of methods

A total of 36 Illumina contigs was considered to be chromosomal. Another one was assumed to contain plasmid-borne sequences (including *blaZ* and *cadX*; see below). The average fragment site of the library was 220 nt. Visual inspection and comparison of these contigs to the Nanopore sequences revealed faulty assemblies of four contigs that needed to be split into two "sub-contigs" each in order to allow alignment to the Nanopore sequence. Most significantly, Illumina/velvet failed to resolve a ca. 5,000 nt region within the ACME-III element that consisted of repetitive sequences. On the other hand, Nanopore showed a poor resolution of poly-A and poly-T sequence fragments resulting in the loss of approximately 15,800 nucleotides across the entire chromosome.

Characterisation of the clinical isolate

Array hybridization revealed the presence of the enterotoxin homologue ORF-CM14 and of an ACME-III element, as well as the absence of *edinB* and *etD*. Otherwise, the isolate matched previously characterized CC80 strains (see Supplemental file 1). In order to explain these discrepancies, it was sequenced using both, Illumina and Nanopore methods and resulting sequences were aligned resulting in a continuous chromosome with a total length of 2,789,663 nt and an overall G/C content of 32.98%. MLST was performed based on the consensus genome sequence and it yielded ST-80 (*arcC*-1, *aroE*-3, *glpF*-1, *gmk*-14, *pta*-11, *tpi*-51, *yqiL*-10).

A comparison of core chromosomal genes revealed that two regions in ANR570100 differed from CP003194. When visually inspecting the mapping of the ANR570100 reads to

CP003194, in those two regions we were able to identify the true extend of these deviant regions (Figure 1).

Deviating Region 1

Deviating Region 1 extends from *walJ* (locus tag MS7_0024 in CC80, CP003194, or respectively SACOL0023 in CC8, CP000046) with a putative recombination sites located in the intergenic region between *walL* and *walJ*. It extends to certainly include *sirB* (MS7_0106, SACOL0098), possibly even to *sbnE* (MS7_0112, SACOL0104) although the differences to canonical CC80 sequences are not large enough to clearly determine a recombination site. It can be estimated at 84,363 nt (based on a consensus of the Illumina and the Nanopore sequences, and including *walJ* to *sirB*). This corresponds to roughly 3% of the genome and includes *ca.* 34,000 nt belonging to the ACME-element.

The gene content of Deviating Region 1 is described in Table 1 where it is also compared to a CC80 reference sequence CP003194 as well as to 02T-671.

Deviating Region 1 consists of four different fragments. The first comprises the genes between *walJ* (MS7_0024; SACOL0023) and *orfX*.

The second one is an ACME-III element including the *opp* operon. This is a potentially motile element and thus it is not necessarily connected to the genomic replacement in this strain. It will be discussed below.

A third fragment includes, among other genes, the enterotoxin homologue ORF-CM14. It extends to Q7A890/Q2YUT2 (MS7_0086/MS7_0087; SACOL0076/SACOL0077). This fragment does not contain the enterotoxin H gene *seh* or a *seh*-derived pseudogene (MS7_0080) which are characteristic for CC1 or CC80, respectively.

A forth fragment of Deviating Region 1 includes the gene encoding staphylococcal protein A. It can be assigned to RIDOM *spa* type t1849; 07-23-34-33-13. This *spa* type is related, but not identical, to both, those of CC1 (such as t127; 07-23-21-16-34-33-13) and CC80 (such as t044; 07-23-12-34-34-33-34). The RIDOM database shows 10 matches (https://spa.ridom.de/spa-t1849.shtml; as of 2020, April 3rd), including three belonging to a German project on characterisation of African *S. aureus* isolates [29, 35, 36] but without disclosing their MLST types or further details. Several other genes in this fragment match CC1 sequences (see Table 1).

The SCC element as part of Deviating Region 1

Deviating Region 1 also comprises *orfX* together with integrated SCC elements. The reference sequence CP003194 contains a SCC*mec* IVc element and most published isolates and sequences of CC80 harbour SCC*mec* IVa or IVc elements. These are absent from ANR570100. Instead, it carries an SCC element without *mecA/C* genes.

The gene content of the SCC element is summarized in Table 2. In short, it consists of

- a type II restriction-modification system,
- *ccrA/B1* recombinase genes and adjacent genes showing some similarity or relationship to SCC*mec* IX sequences (strain JCSC6690, GenBank AB705452.1),
- a large gene with repetitive sequences that is very similar to the gene encoding a hypothetical protein DLJ55_14705 in the chromosomal DNA of strain MOK042 (GenBank CP029627.1) as well as on a plasmid of a ST508/CC45 strain, AR_0471 (chromosome CP029652.1, plasmid CP029650.1)
- and an oligopeptide permease operon, *i.e.*, *opp* genes or ACME-III as well as some genes for “putative proteins” as known from the ST42 strain C427, GenBank ACSQ.

A search of the short read archive of GenBank revealed two near-identical sequences of deviant CC80 strains, one of which (SAMEA48342418) lacked ACME-III while the other one (SAMEA3671725) harboured it, indicating a variable presence of this element in CC80 [ORF-CM14+] strains. When performing a BLAST search with the NCBI GenBank, no significant hits over the entire length of the SCC element were obtained indicating that this element has not yet been observed, although most of its genes have already been found in other SCC elements.

Identification and characterisation of the ST567 isolate 02T-671 as a potential donor of Deviating Region 1

The observation of the enterotoxin homologue ORF-CM14 rather than of the enterotoxin H gene *seh* normally present in canonical CC1 strains, followed by a set of CC1-like genes strongly suggests that Deviating Region 1 is of chimeric origin itself. Our database of ca. 25,000 microarray hybridization profiles was searched for potential donors of Deviating Region 1, *i.e.*, for strains that are chimeras harbouring ORF-CM14 in an otherwise CC1-like core genomic backbone. Only one isolate, 02T-671 a deviant, ST567/CC1 (MLST profile 10-1-1-1-1-1-1, *spa* type, t1242; 07-23-12-34-34-16-34-33-13) strain matched these criteria. However, since no genome sequence was available yet for that strain the isolate that was typed by microarray based-assays was also sequenced using Illumina Miseq.

02T-671 was a PVL-positive CC1 MSSA that differed from canonical CC1 in several features including a presence of the ORF-CM14 enterotoxin homologue and an absence of *seh*. Other differences compared to canonical CC1 strains were the presence of deviant alleles of *aur* and *isaB* as well as an absence of *cna* and Q2G1R6/*cstB* (BA000033.2: 66419-67753). It also harboured an ACME-III element (*opp* genes and *ccrAB1*). The MLST marker *arcC* was different compared to ST1 (*arcC* 10 instead of *arcC* 1) but this difference is due to a single nucleotide polymorphism indicating mutation rather than recombination.

These observations are consistent with integration of a large alien insert around *oriC*. Excluding the ACME-III element, this insert can therefore be estimated to comprise roughly 150,000 nt, ranging from between *arcC* and *aur* across *oriC* and *orfX* to Q7A890/Q2YUT2 (see Figure 1).

Deviating Region 1 of ANR570100 and the corresponding region in the ST567 isolate 02T-671 can be considered identical. This includes the gene content and the gene sequences, the presence and sequence of an ACME-III element and the fault line separating a region of unknown origin from CC1-like sequences between Q7A890 and Q2YUT2.

The ACME-III elements of ANR570100 and 02T-671 were largely identical to each other in both, gene content and gene sequences (see Table 2).

Therefore, we assume the lineage or strain represented by isolate 02T-671 to be the donor of Deviating Region 1 in the lineage of ANR570100. However, the lineage of 02T-671 is itself of chimeric nature comprising a large insert of DNA from a yet unidentified donor into a CC1 genome.

Deviating Region 2

Deviating Region 2 (Table 3, Figure 1) extended from *htsB* (MS7_2199, SACOL2166) to Q8NVB9 (MS7_2323, SACOL2297) or to *ecfA2* (MS7_2242; SACOL2211) having a size of 33,939 to 38,645 nt (1.2 to 1.4% of the genome, which is clearly smaller than the corresponding fragment of the CC80 reference sequence which encompasses as much as 115,604 nt). The reason is that it spans the integration site that in canonical CC80 harbours a motile genomic element comprising of *hsdS*, *hsdM*, *etD*, F3TKB7, *edinB* and F5W4X2 (MS7_2226 to MS7_2231). This element is absent from ANR570100. It is also absent from all CC1 sequences. This region also comprises a gene cluster from *rplQ* (MS7_2243; SACOL2212) to *rpsJ* (MS7_2271; SACOL2240) encoding several ribosomal proteins. These

genes are conserved at a very high degree in all sequences of *S. aureus* so that there was not enough variation for meaningful SNP analysis. However, when BLASTing (https://blast.ncbi.nlm.nih.gov/Blast.cgi) contig 16 (that entirely is a part of Deviating Region 2; the others are on 21 and 19), the five highest scoring matches over the entire length of the query sequence (68,165 nt) are CC1 genomes (with, e.g., 23 nt mismatches and 2 nt gaps for BX571857.1). In general, this region in ANR570100 is more related to CC1 than to canonical CC80 sequences. It also appears to be closer to 02T-671 than to MW2 but given the overall similarity of all sequences concerned, this is hard to assess. The adjacent regions, up- and downstream of Deviating Region 2, are very similar in ANR570100, 02T-671 as well as reference CC1 and CC80 sequences.

The *hla* gene and its neighbouring genes

When comparing the sequence as well as the hybridization profile of ANR570100 to the CC80 reference sequence, the absence of the *hla* gene and its neighbouring genes (A5IS45, Q6GHS5, A5IS47, A6U0Y3, Q2FZB4, *i.e.,* MS7_1116 to MS7_1120 or SACOL1171 to SACOL1175) can be detected. This is presumably the result of homologous recombination between extended repeat sequences flanking the *hla* gene cluster at both of its ends. The presence of *hla* appears to be variable in the deviant CC80 lineage; SAMEA48342418 also lacks *hla* while it is present in SAMEA3671725.

Prophages

When excising the phage sequence (Contig-0007:RC, positions 133,678 to end and Contig-0012 positions 1 to 42,938) and performing a NCBI Blast search, the four best matches, with identities of 99.97%, are all PVL phages from CC80 strains, phiSa2wa_st80 (MG029515.1), NCTC13435 (LN831036.1), GR2 (CP010402.1) and 11819-97 (CP003194.1).

The PVL prophage in ANR570100 is integrated into the same site of the chromosome as the one in 11819-97 (CP003194.1). The prophage sequences from both strains are co-linear and they comprise the same set of genes.

The *hlb*-converting phage in ANR570100 differed from CP003194.1, although the virulence-associated genes it carried (*scn*, *sak*) were the same. A NCBI Blast of Contig-0009RC positions 81,223 to 123,488 (as of August 2019) yielded as best matches (with more than 99.6% identity) the *hlb*-converting phages from BB155 (LN854556.1) and 55-99-44 (CP024998.1) and SA17_S6 (CP010941.1). They all belong to ST152 (https://pubmlst.org/bigsdb?db=pubmlst_saureus_seqdef&page=sequenceQuery). This might be attributed to a co-existence and co-evolution of the deviant CC80 lineage and of CC152 in the same geographic region, as the latter CC is known to be predominant at least in parts of Africa [38-44].

Resistance genes

ANR570100 carried the *blaZ/I/R* operon and a cadmium resistance operon *cadD/cadX*, presumably on a plasmid. Genes *aphA3* and *sat* (neomycin and streptothricin resistance) as well as *far1/fusB* and *tet*(K) that frequently can be encountered in canonical CC80, being situated within SCC*mec* elements or on plasmids, were absent.

Discussion

We identified a virulent, PVL-positive CC80 MSSA which differed in several key features from canonical CC80 strains and sequences. Its analysis was performed using three different methods, array hybridisation, Illumina and Nanopore sequencing. While array hybridisation yields less information than sequencing, it can be routinely performed automatically and economically on high numbers of clinical isolates that, in the present case, allowed the identification of the initial isolate ANR570100 as being of special interest as well as of 02T671 as putative donor. Illumina sequencing provided short reads of high quality sequences, but it has difficulties with repetitive sequences which, as the most relevant problem in the current project, led to a virtual miss of DLJ55_14705 within the ACME-III element. Nanopore sequencing proved unreliable with regard especially to poly-T and poly-A sequences, but it can handle repetitive sequences much better which in *S. aureus* also include MSCRAMM genes such as *spa*.

Differences of the target strain to reference sequences of CC80 include two large inserts of DNA from other *S. aureus* lineages, both combined accounting for about 8% of the genome of the strain. While one was located close to *oriC* which appears to be a hotspot for chromosomal replacements (see Introduction), the other one was localised at a distant position. The mechanism for these gene transfers is yet unknown. With two large replacements being present in one single isolate, we assume that such-large scale horizontal gene transfers might be more common in *S. aureus* than previously perceived, and that the resolution of MLST with seven markers is not high enough to identify all chimeric strains. However, the combination and interaction of microarray-based assays and NGS allows the reliable identification of such strains [37].

The most striking features of ANR570100, however, are large regions in its genome that clearly differ from other CC80 sequences. As described above, Deviating Region 1 in the

isolate ANR570100 comprises sequences identical to the ones from the atypical CC1/ST567 strain 02T671. This includes an ACME-III element. It also includes a stretch of DNA upstream and downstream of ACME-III with the latter part including ORF-CM14. Theoretically, this might give a hint on the putative donor of Deviating Region 1.

Possible donors for ORF-CM14 to both, 570100 and 02T671 obviously must include strains form ORF-CM14 positive lineages that are ST12, ST71, ST93, ST121, ST509, ST567, CC772, CC705, ST707, ST760, ST816, ST848, ST1094, ST1643, ST2272, ST2425, ST2616 and ST2972 (based on published sequences and author´s own microarray data). Unfortunately, genome sequences of several of these STs are not available and those that are available do not match fully the sequence of Deviating Region 1. When comparing ORF-CM14 sequences alone, those of JKD6159; CP002114.2 [76914-77693] (ST93) and SS-015; FQIU01000002 [597790-598569] (ST2972) are the most closely related ones. When performing BLAST on the non-CC1-region of 02T671, the highest scoring hits are two ST2272 sequences (AP019712.1 and AP019713.1). When directly comparing sequences in question, the differences are large enough to indicate that ST2272 was not likely to be the direct donor (with an average difference of 1.8% for *dnaA, dnaN, yaaA, recF, gyrB, gyrA, nnrD, hutH, serS, azlC, sam-L1, metX, yybS, gdpP, rplI, dnaC, purA, walR, walK, walH, walI, walJ, sasH*, Q6GKL1, Q6GKL6, ORF-CM14*, dusC*, A6TXM6, A6QD71, Q6GKK6 and Q2YUT2 from Tokyo12482, GenBank AP019713.1, to 02T671).

In both strains, ANR570100 and 02T671, a fault-line can be observed between Q7A890 and Q2YUT2 separating downstream sequences of unknown origin from those upstream that are rather unambiguously related to CC1 (*i.e.*, the right border between “red” and “blue” sectors in Figure 1 and the last two columns of Table 1). This means Deviating Region 1 of ANR570100 includes the fault line separating the alien insert in 02T671 from the canonical CC1 core genome of that strain. This makes it very likely that a 02T671-like strain was indeed

the donor of Deviating Region 1, and that this region itself is of chimeric nature, spanning CC1 and non-CC1 sequences that together form the 02T671-like donor strain as well as the mobile SCC/ACME-III element (see Figure 1). Unfortunately, the upstream fault line separating CC1 from CC80 sequences in ANR570100 (between *sirB* and *spa* or *sbnE*) cannot exactly been determined because of the general similarity or relatedness of CC1 and CC80.

Deviating Region 1 also comprises an apparently new ACME-III element. The presence of *opp* genes and *ccrA/B-1* recombinase genes are reminiscent of the CC34 strain 21342 (GenBank AHKU) although the sequence of *ccrB*-1 appears to be more related to the one from SCC*mec* IX. It also includes, as revealed mainly by Nanopore sequencing, a gene with repetitive sequences that is very similar to the gene encoding a hypothetical protein DLJ55_14705 in strain MOK042. This strain belongs to ST71, a lineage that also can be described as chimera, comprising of a large insert of unknown origin in a CC97 genome. In strain MOK042, the gene encoding DLJ55_14705 is localised on that insert but it is not a part of a SCC element.

In addition, there is a second Deviant Region elsewhere in the genome of the lineage of ANR570100. Its gene content as well as its gene sequences are highly similar to 02T-671 and to reference CC1 sequences but they clearly differ from canonical CC80. These differences include, but are not limited to, the absence of *edinB* and *etD*. Unfortunately, the region in question includes genes whose origin cannot be determined because of a high degree of conservation of the genes affected. For the same reason, the exact boundaries of the Deviating Region cannot be identified. Interestingly, the adjacent regions to Deviant Region 2 are very similar in all sequences analysed, *i.e.*, ANR570100, 02T-671 as well as the reference CC1 and CC80 sequences (with differences being less than 0.5%). This could suggest that the region corresponding to Deviant Region 2 was “deviant” not in ANR570100 but, compared to the other three sequences, in the CC80 reference sequence. This might indicate that Deviant

373 Region 2 in ANR570100 was not an alien insert of CC1 origin but that its sequence represents
374 shared, ancestral CC1/CC80 stock and that the corresponding region in canonical CC80
375 (including *edinB* and *etD*) itself was an insert from another, yet unidentified, lineage.

376 In conclusion, the core genome of ANR570100 bears evidence of at least two, possibly three
377 large-scale recombination events. First, ORF-CM14, among other genes, was introduced into
378 a CC1 strain and, second, the resulting ORF-CM14/CC1 composite fragment was introduced
379 into CC80. In addition, another recombination event introduced either Deviating Region 2
380 from CC1 into the ancestor of ANR570100 or the corresponding region, possibly together
381 with *edinB* and *etD,* from an unknown donor into canonical CC80.

382 Thus, such complex and large-scale recombination events cannot be that rare and exceptional,
383 despite a mainly clonal nature of *S. aureus* [45]. Although the exact mechanism is not clear,
384 chimerism seems to be an additional pathway in the evolution of *S. aureus*, possibly being
385 responsible for a transmission of virulence factors (such as ORF-CM14 in the case described
386 herein) or of resistance genes including entire SCC*mec* elements [3]. From a more theoretical
387 point of view, large-scale genomic substitutions, chimerism or hybridisation facilitate
388 evolutionary leaps that cannot be achieved by accumulation of single point mutations or that
389 would require immeasurably much more time to be achieved by mutations. If one considers
390 the ability to evolve and adapt as an evolutionary advantage, an organism that can shuffle,
391 swap or exchange major parts of its genome by whatever unknown mechanism should be in a
392 better position than a strictly clonal organism.

393

Acknowledgements

A part of this work was presented during a poster presentation at the 18th International Symposium on Staphylococci and Staphylococcal Infections (ISSSI, 23 – 26 August, 2018 Copenhagen, Denmark).

We thank for the excellent technical assistance of Byrgit Hofmann, Friedrich-Loeffler-Institut, Jena, Germany, as well as Peter Slickers, Jena, for help and advice regarding sequence analyses.

Authors´ contributions

A. Ruppelt-Lorz and B. Söderquist found and identified strains of interest. P. Slickers, R. Ehricht and S. Monecke designed the study. A. Ruppelt-Lorz, E. Müller, D. Gawlik and A. Reißig carried out experiments; S. Monecke analysed the sequence data; H. Hotzel, S. Braun and B. Söderquist performed/supervised sequencing; A. Ziegler-Cordts created a software tool used for sequence analysis; D. Gawlik, R. Ehricht and S. Monecke wrote the manuscript. All authors read and approved the final manuscript.

Funding

There was no external funding for this study.

Competing interests

DG is employee of PTC - Phage Technology Center GmbH, Bönen, Germany; AZC is employee of T-Systems Multimedia Solutions GmbH, Dresden, Germany. In both cases, work on this project was performed before the respective employments started. Thus, employers of the authors did not have any role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The other authors declare that no competing interests exist.

The authors adhere to PLOS ONE policies on sharing data and materials. The specific roles of authors are articulated in the 'author contributions' section.

Availability of data and materials

The genome sequences of ANR570100 and 02T-671 are published in GenBank under the following accession numbers: *submission pending*

Ethics approval and consent to participate

Not applicable.

References

1. Wan TW, Khokhlova OE, Iwao Y, Higuchi W, Hung WC, Reva IV, et al. Complete Circular Genome Sequence of Successful ST8/SCC*mec*IV Community-Associated Methicillin-Resistant *Staphylococcus aureus* (OC8) in Russia: One-Megabase Genomic Inversion, IS256's Spread, and Evolution of Russia ST8-IV. PLoS ONE. 2016;11(10):e0164168. Epub 2016/10/16. doi: 10.1371/journal.pone.0164168. PubMed PMID: 27741255; PubMed Central PMCID: PMC5065196.

2. Robinson DA, Enright MC. Evolution of *Staphylococcus aureus* by Large Chromosomal Replacements. J Bacteriol. 2004;186(4):1060-4.

3. Nimmo GR, Steen JA, Monecke S, Ehricht R, Slickers P, Thomas JC, et al. ST2249-MRSA-III: a second major recombinant methicillin-resistant *Staphylococcus aureus* clone causing healthcare infection in the 1970s. Clin Microbiol Infect. 2015;21(5):444-50. Epub 2015/02/25. doi: 10.1016/j.cmi.2014.12.018. PubMed PMID: 25708549; PubMed Central PMCID: PMC4564996.

4. Fetsch A, Kraushaar B, Kasbohrer A, Hammerl JA. Turkey Meat as Source of CC9/CC398 Methicillin-Resistant *Staphylococcus aureus* in Humans? Clin Infect Dis. 2017;64(1):102-3. Epub 2016/10/30. doi: 10.1093/cid/ciw687. PubMed PMID: 27794020; PubMed Central PMCID: PMC5159606.

5. Monecke S, Slickers P, Gawlik D, Müller E, Reissig A, Ruppelt-Lorz A, et al. Variability of SCC*mec* elements in livestock-associated CC398 MRSA. Veterinary Microbiology. 2018;217:36-46. doi: 10.1016/j.vetmic.2018.02.024.

6. Faria NA, Oliveira DC, Westh H, Monnet DL, Larsen AR, Skov R, et al. Epidemiology of emerging methicillin-resistant *Staphylococcus aureus* (MRSA) in Denmark: a nationwide study in a country with low prevalence of MRSA infection. J Clin Microbiol. 2005;43(4):1836-42. PubMed PMID: 15815005.

7. Holmes A, Ganner M, McGuane S, Pitt TL, Cookson BD, Kearns AM. *Staphylococcus aureus* isolates carrying Panton-Valentine leucocidin genes in England and Wales: frequency, characterization, and association with clinical disease. J Clin Microbiol. 2005;43(5):2384-90. PubMed PMID: 15872271.

8. Krziwanek K, Luger C, Sammer B, Stumvoll S, Stammler M, Metz-Gercek S, et al. PVL-positive MRSA in Austria. Eur J Clin Microbiol Infect Dis. 2007;26(12):931-5. PubMed PMID: 17891548.

9. Monecke S, Aamot HV, Stieber B, Ruppelt A, Ehricht R. Characterization of PVL-positive MRSA from Norway. Apmis. 2014;122(7):580-4. Epub 2013/10/11. doi: 10.1111/apm.12181. PubMed PMID: 24106794.

10. Monecke S, Slickers P, Hotzel H, Richter-Huhn G, Pohle M, Weber S, et al. Microarray-based characterisation of a Panton-Valentine leukocidin-positive community-acquired strain of methicillin-resistant *Staphylococcus aureus*. Clin Microbiol Infect. 2006;12(8):718-28.

11. Monecke S, Jatzwauk L, Weber S, Slickers P, Ehricht R. DNA microarray-based genotyping of methicillin-resistant *Staphylococcus aureus* strains from Eastern Saxony. Clin Microbiol Infect. 2008;14(6):534-45. Epub 2008/04/01. doi: 10.1111/j.1469-0691.2008.01986.x. PubMed PMID: 18373691.

12. Otter JA, French GL. The emergence of community-associated methicillin-resistant *Staphylococcus aureus* at a London teaching hospital, 2000-2006. Clin Microbiol Infect. 2008;14(7):670-6. PubMed PMID: 18558939.

13. Rossney AS, Shore AC, Morgan PM, Fitzgibbon MM, O'Connell B, Coleman DC. The emergence and importation of diverse genotypes of methicillin-resistant *Staphylococcus aureus* (MRSA) harboring the Panton-Valentine leukocidin gene (pvl) reveal that pvl is a poor marker for community-acquired MRSA strains in Ireland. J Clin Microbiol. 2007;45(8):2554-63. PubMed PMID: 17581935.

14. Stegger M, Wirth T, Andersen PS, Skov RL, De Grassi A, Simoes PM, et al. Origin and evolution of European community-acquired methicillin-resistant *Staphylococcus aureus*. mbio. 2014;5(5):e01044-14. Epub 2014/08/28. doi: 10.1128/mBio.01044-14. PubMed PMID: 25161186; PubMed Central PMCID: PMC4173770.

15. Aires de Sousa M, Bartzavali C, Spiliopoulou I, Sanches IS, Crisostomo MI, de Lencastre H. Two international methicillin-resistant *Staphylococcus aureus* clones endemic in a university hospital in Patras, Greece. J Clin Microbiol. 2003;41(5):2027-32. PubMed PMID: 12734244.

16. Vourli S, Perimeni D, Makri A, Polemis M, Voyiatzi A, Vatopoulos A. Community acquired MRSA infections in a paediatric population in Greece. Euro Surveill. 2005;10(5):78-9. PubMed PMID: 16077207.

17. Maier J, Melzl H, Reischl U, Drubel I, Witte W, Lehn N, et al. Panton-Valentine leukocidin-positive methicillin-resistant *Staphylococcus aureus* in Germany associated with travel or foreign family origin. Eur J Clin Microbiol Infect Dis. 2005;24(9):637-9.

18. Tokajian ST, Khalil PA, Jabbour D, Rizk M, Farah MJ, Hashwa FA, et al. Molecular characterization of *Staphylococcus aureus* in Lebanon. Epidemiol Infect. 2010;138(5):707-12. PubMed PMID: 20202283.

19. Scicluna E, Shore A, Thuermer A, Ehricht R, Slickers P, Borg M, et al. Characterisation of MRSA from Malta and the description of a Maltese epidemic MRSA strain. Eur J Clin Microbiol Infect Dis. 2010;29(2):163-70.

20. Ben Nejma M, Mastouri M, Bel Hadj Jrad B, Nour M. Characterization of ST80 Panton-Valentine leukocidin-positive community-acquired methicillin-resistant *Staphylococcus aureus* clone in Tunisia. Diagn Microbiol Infect Dis. 2013;77(1):20-4. Epub 2008/04/09. doi: 10.1016/j.diagmicrobio.2008.02.010. PubMed PMID: 18394845.

21. Ben Slama K, Gharsa H, Klibi N, Jouini A, Lozano C, Gomez-Sanz E, et al. Nasal carriage of *Staphylococcus aureus* in healthy humans with different levels of contact with animals in Tunisia: genetic lineages, methicillin resistance, and virulence factors. Eur J Clin Microbiol Infect Dis. 2011;30(4):499-508. Epub 2010/11/16. doi: 10.1007/s10096-010-1109-6. PubMed PMID: 21076928.

22. Antri K, Rouzic N, Dauwalder O, Boubekri I, Bes M, Lina G, et al. High prevalence of methicillin-resistant *Staphylococcus aureus* clone ST80-IV in hospital and community settings in Algiers. Clin Microbiol Infect. 2011;17(4):526-32.

23. Basset P, Amhis W, Blanc DS. Changing molecular epidemiology of methicillin-resistant *Staphylococcus aureus* in an Algerian hospital. J Infect Dev Ctries. 2015;9(2):206-9. Epub 2015/02/24. doi: 10.3855/jidc.4620. PubMed PMID: 25699496.

24. Djoudi F, Bonura C, Benallaoua S, Touati A, Touati D, Aleo A, et al. Panton-Valentine leukocidin positive sequence type 80 methicillin-resistant *Staphylococcus aureus* carrying a staphylococcal cassette chromosome *mec* type IVc is dominant in neonates and children in an Algiers hospital. New Microbiol. 2013;36(1):49-55. Epub 2013/02/26. PubMed PMID: 23435815.

25. Boswihi SS, Udo EE, Al-Sweih N. Shifts in the Clonal Distribution of Methicillin-Resistant *Staphylococcus aureus* in Kuwait Hospitals: 1992-2010. PLoS ONE. 2016;11(9):e0162744. doi: 10.1371/journal.pone.0162744.

26. Monecke S, Skakni L, Hasan R, Ruppelt A, Ghazal SS, Hakawi A, et al. Characterisation of MRSA strains isolated from patients in a hospital in Riyadh, Kingdom of Saudi Arabia. BMC Microbiol. 2012;12(1):146. Epub 2012/07/25. doi: 10.1186/1471-2180-12-146. PubMed PMID: 22823982.

27. Udo EE, Al-Lawati BAH, Al-Muharmi Z, Thukral SS. Genotyping of methicillin-resistant *Staphylococcus aureus* in the Sultan Qaboos University Hospital, Oman reveals the dominance of Panton–Valentine leucocidin-negative ST6-IV/t304 clone. New Microbes and New Infections. 2014;2(4):100-5. doi: 10.1002/nmi2.47. PubMed PMID: PMC4184578.

28. Udo EE, Sarkhoo E. Genetic analysis of high-level mupirocin resistance in the ST80 clone of community-associated meticillin-resistant *Staphylococcus aureus*. J Med Microbiol. 2010;59(Pt 2):193-9. PubMed PMID: 19833783.

29. Shittu AO, Oyedara O, Kenneth OO, Raji A, Peters G, von Müller L, et al. An assessment on DNA microarray and sequence-based methods for the characterization of methicillin-susceptible *Staphylococcus aureus* from Nigeria. Frontiers in Microbiology. 2015;6. doi: 10.3389/fmicb.2015.01160.

30. Ghebremedhin B, Olugbosi MO, Raji AM, Layer F, Bakare RA, Konig B, et al. Emergence of a community-associated methicillin-resistant *Staphylococcus aureus* strain with a unique resistance profile in Southwest Nigeria. J Clin Microbiol. 2009;47(9):2975-80. PubMed PMID: 19571020.

31. Schaumburg F, Kock R, Friedrich AW, Soulanoudjingar S, Ngoa UA, von Eiff C, et al. Population structure of *Staphylococcus aureus* from remote African Babongo Pygmies. PLoS Negl Trop Dis. 2011;5(5):e1150. Epub 2011/05/17. doi: 10.1371/journal.pntd.0001150. PubMed PMID: 21572985; PubMed Central PMCID: PMC3091839.

32. Monecke S, Coombs G, Shore AC, Coleman DC, Akpaka P, Borg M, et al. A field guide to pandemic, epidemic and sporadic clones of methicillin-resistant *Staphylococcus aureus*. PLoS One. 2011;6(4):e17936. doi: 10.1371/journal.pone.0017936.

33. Monecke S, Jatzwauk L, Muller E, Nitschke H, Pfohl K, Slickers P, et al. Diversity of SCC*mec* elements in *Staphylococcus aureus* as observed in South-Eastern Germany. PLoS ONE. 2016;11(9):e0162654. Epub 2016/09/21. doi: 10.1371/journal.pone.0162654. PubMed PMID: 27648947.

34. Monecke S, Slickers P, Ehricht R. Assignment of *Staphylococcus aureus* isolates to clonal complexes based on microarray analysis and pattern recognition. FEMS Immunol Med Microbiol. 2008;53:237–51.

35. Herrmann M, Abdullah S, Alabi A, Alonso P, Friedrich AW, Fuhr G, et al. Staphylococcal disease in Africa: another neglected 'tropical' disease. Future Microbiol. 2013;8(1):17-26. Epub 2012/12/21. doi: 10.2217/fmb.12.126. PubMed PMID: 23252490.

36. Strauß L, Ruffing U, Abdulla S, Alabi A, Akulenko R, Garrine M, et al. Detecting *Staphylococcus aureus* Virulence and Resistance Genes: a Comparison of Whole-Genome Sequencing and DNA Microarray Technology. Journal of Clinical Microbiology. 2016;54(4):1008-16. doi: 10.1128/jcm.03022-15.

37. Dunne WM, Pouseele H, Monecke S, Ehricht R, van Belkum A. Epidemiology of transmissible diseases: Array hybridization and next generation sequencing as universal nucleic acid-mediated typing tools. Infection Genetics and Evolution. 2018;63:332-45. doi: 10.1016/j.meegid.2017.09.019. PubMed PMID: WOS:000442167600043.

38. Conceicao T, Coelho C, Silva IS, de Lencastre H, Aires-de-Sousa M. *Staphylococcus aureus* in former Portuguese colonies from Africa and the Far East: missing data to help fill the world map. Clin Microbiol Infect. 2015;21(9):842 e1- e10. Epub 2015/05/25. doi: 10.1016/j.cmi.2015.05.010. PubMed PMID: 26003281.

39. Conceicao T, Santos Silva I, de Lencastre H, Aires-de-Sousa M. *Staphylococcus aureus* nasal carriage among patients and health care workers in Sao Tome and Principe. Microb Drug Resist. 2014;20(1):57-66. Epub 2013/09/13. doi: 10.1089/mdr.2013.0136. PubMed PMID: 24024594.

40. Egyir B, Guardabassi L, Sorum M, Nielsen SS, Kolekang A, Frimpong E, et al. Molecular epidemiology and antimicrobial susceptibility of clinical *Staphylococcus aureus* from healthcare institutions in Ghana. PLoS ONE. 2014;9(2):e89716. Epub 2014/03/04. doi: 10.1371/journal.pone.0089716. PubMed PMID: 24586981; PubMed Central PMCID: PMC3934920.

41. Ruimy R, Maiga A, Armand-Lefevre L, Maiga I, Diallo A, Koumare AK, et al. The carriage population of *Staphylococcus aureus* from Mali is composed of a combination of pandemic clones and the divergent Panton-Valentine leukocidin-positive genotype ST152. J Bacteriol. 2008;190(11):3962-8. PubMed PMID: 18375551.

42. Shittu A, Oyedara O, Abegunrin F, Okon K, Raji A, Taiwo S, et al. Characterization of methicillin-susceptible and -resistant staphylococci in the clinical setting: a multicentre study in Nigeria. BMC Infect Dis. 2012;12:286. Epub 2012/11/06. doi: 10.1186/1471-2334-12-286. PubMed PMID: 23121720; PubMed Central PMCID: PMC3529121.

43. Shittu AO, Okon K, Adesida S, Oyedara O, Witte W, Strommenger B, et al. Antibiotic resistance and molecular epidemiology of *Staphylococcus aureus* in Nigeria. BMC Microbiol. 2011;11:92. PubMed PMID: 21545717.

44. Okuda KV, Toepfner N, Alabi AS, Arnold B, Belard S, Falke U, et al. Molecular epidemiology of *Staphylococcus aureus* from Lambarene, Gabon. Eur J Clin Microbiol Infect Dis. 2016. doi: 10.1007/s10096-016-2748-z. PubMed PMID: 27553495.

45. Feil EJ, Cooper JE, Grundmann H, Robinson DA, Enright MC, Berendt T, et al. How clonal is *Staphylococcus aureus*? J Bacteriol. 2003;185(11):3307-16. PubMed PMID: 12754228.

Tables, Figures and Supplemental Files

Table 1: Genes in Deviating Region 1 in comparison to canonical CC80, to 02T-671 and to canonical CC1.

Table 2: The ACME-III element in ANR570100 and 02T-671.

Table 3: Genes in Deviating Region 2 in comparison to canonical CC80, to 02T-671 and to canonical CC1.

Figure 1: Schematic diagram of the genomes of 02T567 (outer circle), ANR570100 (middle circle) and the reference genome CC80, 11819-97, GenBank CP003194.1 (inner circle). Genomic fragments are colour-coded depending on their origin.

Supplemental file 1: Array hybridization patterns of strains discussed.

Supplemental file 2A: Illumina and Nanopore Consensus sequence of ANR570100.

Supplemental file 2B: Annotated sequence of ANR570100.

Supplemental file 3A: Illumina sequence of 02T671.

Supplemental file 3B: Annotated sequence of 02T671.

Supplemental file 4: fasta-file with the ACME III sequences and markers.

Table 1: Genes in Deviating Region 1 in comparison to canonical CC80, to 02T-671 and to canonical CC1.

	Comparison of ANR570100 to CC80 reference sequence 11819-97 (CP003194)	**Comparison of ANR570100 to 02T-671**	**Comparison of 02T-671 to MW2**	**Provenance in ANR570100**	**Provenance in 02T-671**
walR	0 mismatches/702 bases (0%)	2 mismatches/702 bases (0.28%)	4 mismatches/702 bases (0.57%)	CC80	unknown
walK	5 mismatches/1827 bases (0.27%)	12 mismatches/1827 bases (0.66%)	14 mismatches/1827 bases (0.77%)	CC80	unknown
walH	18 mismatches/1335 bases (1.35%)	36 mismatches/1335 bases (2.7%)	36 mismatches/1335 bases (2.7%)	CC80	unknown
walI	0 mismatches/789 bases (0%)	19 mismatches/789 bases (2.41%)	19 mismatches/789 bases (2.41%)	CC80	unknown
walJ	17 mismatches/801 bases (2.12%)	0 mismatches/801 bases (0%)	18 mismatches/801 bases (2.25%)	unknown	unknown
sasH	95 mismatches/2319 bases (4.1%)	0 mismatches/2319 bases (0%)	65 mismatches/2319 bases (2.8%)	unknown	unknown
orfX	16 mismatches/480 bases (3.33%)	0 mismatches/480 bases (0%)	14 mismatches/480 bases (2.92%)	unknown	unknown
ACME-III (see Table 2*)*	*Not present in 11819-97 (which has SCCmec IVc instead)*	Identical element (see text and Table 3)	*Not present in MW2 (which has SCCmec IVa instead)*	ACME-III	ACME-III
Q6GKL1	*Not present in 11819-97/in canonical CC80*	0 mismatches/576 bases (0%)	*Not present in MW2/in canonical CC1*	unknown	unknown
Q6GKL6	*Not present in 11819-97/in canonical CC80*	0 mismatches/159 bases (0%)	*Not present in MW2/in canonical CC1*	unknown	unknown
ORF-CM14	*Not present in 11819-97/in canonical CC80*	0 mismatches/780 bases (0%)	*Not present in MW2/in canonical CC1*	unknown	unknown
dusC	29 mismatches/987 bases (2.94%)	0 mismatches/987 bases (0%)	26 mismatches/987 bases (2.63%)	unknown	unknown
A6TXM6	23 mismatches/204 bases (11.27%)	0 mismatches/204 bases (0%)	25 mismatches/204 bases (12.25%)	unknown	unknown
A6QD71	31 mismatches/297 bases (10.44%)	0 mismatches/297 bases (0%)	32 mismatches/297 bases (10.77%)	unknown	unknown
Q5HJT2	*Absent from ANR570100 and canonical CC80*	*Absent from ANR570100 and 02T-671*	*Absent from 02T-671 but present in canonical CC1*	unknown	unknown
Q6GD34	*Absent from ANR570100 and canonical CC80*	*Absent from ANR570100 and 02T-671*	*Absent from 02T-671 but present in canonical CC1*	unknown	unknown
A6QD75	*Absent from ANR570100 and canonical CC80*	*Absent from ANR570100 and 02T-671*	*Absent from 02T-671 but present in canonical CC1*	unknown	unknown
A6QD76	*Absent from ANR570100 and canonical CC80*	*Absent from ANR570100 and 02T-671*	*Absent from 02T-671 but present in canonical CC1*	unknown	unknown
A8YZ18	*Absent from ANR570100 and canonical CC80*	*Absent from ANR570100 and 02T-671*	*Absent from 02T-671 but present in canonical CC1*	unknown	unknown
Q6GKK6	22 mismatches/612 bases (3.6%)	1 mismatches/612 bases (0.16%)	17 mismatches/612 bases (2.78%)	unknown	unknown
Q7A890	*Not present in 11819-97*	1 mismatches/3153 bases (0.03%)	104 mismatches/3153 bases (3.3%)	unknown	unknown
Q2YUT2	1 mismatches/483 bases (0.21%)	0 mismatches/483 bases (0%)	2 mismatches/483 bases (0.41%)	CC1 or CC80	CC1
plc	26 mismatches/987 bases (2.63%)	0 mismatches/987 bases (0%)	1 mismatches/987 bases (0.1%)	CC1	CC1
***lpl*-SAOUHSC_00052**	17 mismatches/771 bases (2.2%)	0 mismatches/771 bases (0%)	1 mismatches/771 bases (0.13%)	CC1	CC1
***lpl*-SAOUHSC_00053**	97 mismatches/771 bases (12.58%)	0 mismatches/771 bases (0%)	1 mismatches/771 bases (0.13%)	CC1	CC1
***lpl*-MW0073**	100 mismatches/693 bases (14.43%)	0 mismatches/693 bases (0%)	0 mismatches/693 bases (0%)	CC1	CC1
***lipC3*-MW0074**	42 mismatches/1377 bases (3.05%)	0 mismatches/1377 bases (0%)	0 mismatches/1377 bases (0%)	CC1	CC1
Q8NYT6	17 mismatches/2238 bases (0.76%)	1 mismatches/2238 bases (0.04%)	20 mismatches/2238 bases (0.89%)	CC1	CC1 ?
Teg15as	0 mismatches/244 bases (0%)	0 mismatches/244 bases (0%)	0 mismatches/244 bases (0%)	*Related in all lineages in question*	*Related in all lineages in question*

	Comparison of ANR570100 to CC80 reference sequence 11819-97 (CP003194)	**Comparison of ANR570100 to 02T-671**	**Comparison of 02T-671 to MW2**	**Provenance in ANR570100**	**Provenance in 02T-671**
Q8NYT5	23 mismatches/1179 bases (1.95%)	0 mismatches/1179 bases (0%)	1 mismatches/1179 bases (0.08%)	CC1	CC1
norC	27 mismatches/1389 bases (1.94%)	0 mismatches/1389 bases (0%)	0 mismatches/1389 bases (0%)	CC1	CC1
nptA	14 mismatches/1662 bases (0.84%)	0 mismatches/1662 bases (0%)	0 mismatches/1389 bases (0%)	CC1	CC1
Q2YUS5	17 mismatches/1776 bases (0.96%)	0 mismatches/1776 bases (0%)	1 mismatches/1776 bases (0.06%)	CC1	CC1
DUF1648	0 mismatches/474 bases (0%)	0 mismatches/474 bases (0%)	0 mismatches/474 bases (0%)	*Related in all lineages in question*	*Related in all lineages in question*
***lctP*-locus1**	20 mismatches/1593 bases (1.26%)	0 mismatches/1593 bases (0%)	0 mismatches/1593 bases (0%)	CC1	CC1
txbi_lctP	1 mismatches/72 bases (1.39%)	0 mismatches/72 bases (0%)	0 mismatches/72 bases (0%)	CC1	CC1
txbi_proteinA	1 mismatches/62 bases (1.61%)	0 mismatches/62 bases (0%)	0 mismatches/62 bases (0%)	CC1	CC1
spa	ANR570100: *spa* t 1849 **07-23-34-33-13** 11819-97: *spa* t044 **07-23-12-34-34-33-34**	ANR570100: *spa* t 1849 **07-23-34-33-13** 02T671: *spa* t1242 **07-23-12-34-34-16-34-33-13**	02T671: *spa* t1242 **07-23-12-34-34-16-34-33-13** MW2: *spa* t128 **07-23-23-21-16-34-33-13**	*Related in all lineages in question*	*Related in all lineages in question*
tx_sarS	0 mismatches/63 bases (0%)	0 mismatches/63 bases (0%)	0 mismatches/63 bases (0%)	*Related in all lineages in question*	*Related in all lineages in question*
sarS	5 mismatches/753 bases (0.66%)	0 mismatches/753 bases (0%)	0 mismatches/753 bases (0%)	CC1	CC1
sirC	7 mismatches/999 bases (0.7%)	0 mismatches/999 bases (0%)	0 mismatches/999 bases (0%)	CC1	CC1
sirB	4 mismatches/996 bases (0.4%)	5 mismatches/996 bases (0.5%)	0 mismatches/996 bases (0%)	CC1 or CC80	CC1
sirA	0 mismatches/993 bases (0%)	6 mismatches/993 bases (0.6%)	9 mismatches/993 bases (0.91%)	CC80	CC1 ?
sbnA	0 mismatches/981 bases (0%)	1 mismatches/981 bases (0.1%)	0 mismatches/981 bases (0%)	*Related in all lineages in question*	*Related in all lineages in question*
sbnB	1 mismatches/1011 bases (0.1%)	2 mismatches/1011 bases (0.2%)	1 mismatches/1011 bases (0.1%)	CC1 or CC80	CC1
sbnC	4 mismatches/1755 bases (0.23%)	9 mismatches/1755 bases (0.51%)	0 mismatches/1755 bases (0%)	CC80	CC1
sbnD	3 mismatches/1257 bases (0.24%)	0 mismatches/1257 bases (0%)	3 mismatches/1257 bases (0.24%)	CC1 or CC80	CC1
sbnE	16 mismatches/1737 bases (0.92%)	17 mismatches/1737 bases (0.98%)	8 mismatches/1737 bases (0.46%)	CC1 or CC80	CC1
sbnF	0 mismatches/1740 bases (0%)	8 mismatches/1740 bases (0.46%)	13 mismatches/1740 bases (0.75%)	CC80	CC1 ?
sbnG	4 mismatches/777 bases (0.51%)	4 mismatches/777 bases (0.51%)	1 mismatches/777 bases (0.13%)	CC1 or CC80	CC1
sbnH	5 mismatches/1203 bases (0.42%)	3 mismatches/1203 bases (0.25%)	2 mismatches/1203 bases (0.17%)	CC1 or CC80	CC1
sbnI	3 mismatches/765 bases (0.39%)	4 mismatches/765 bases (0.52%)	2 mismatches/765 bases (0.26%)	CC80	CC1
Q5HJP4	*Absent from ANR570100 and 11819-97*	*Absent from ANR570100 but present in 02T-671*	1 mismatches/408 bases (0.25%)	CC80	CC1

Table 2: The ACME-III element in ANR570100 and 02T-671.

Gene	Description/gene product and comments	Orientation	Start position in SCC	End position in SCC	Start position in genome	End position in genome	Comparison of ANR570100 to 02T-671
orfX	23S rRNA methyltransferase with the SCC integration site being located at the 3' end of *orfX*.	Forward	1	480	33682	34162	0 mismatches/480 bases (0%)
sRNA6	Antisense RNA associated with *orfX*.		181	464	33862	34146	0 mismatches/284 bases (0%)
DR_SCC	Direct repeat of SCC, 19 nt of the 3' end of the coding sequence of *orfX*.		462	480	34143	34162	0 mismatches/19 bases (0%)
dam5	Type II restriction-modification system, endonuclease and methyltransferase. A reference sequence for this gene is from strain K12S0375, GenBank JYGF01000026.1 [127750:130506].	Forward	775	3522	34456	37204	0 mismatches/2748 bases (0%)
helicase	DNA helicase, associated with *dam,* putative restriction system. A reference sequence for this gene is from strain K12S0375, GenBank JYGF01000026.1 [130502:132466].	Forward	3512	5477	37193	39159	0 mismatches/1966 bases (0%)
"YeeC"	YeeC-like protein. 1381/1442(96%) identities and 4/1442 gaps compared to strain FORC_090, GenBank CP029198.1:39824-41262 (FORC090_0030)	Reverse	5480	6920	39161	40602	0 mismatches/1441 bases (0%)
A9UFT0	LPXTG protein homologue. A reference sequence for this gene is from strain C427-ST42, GenBank ACSQ01000048.1 [74:295].	Reverse	6685	6906	40366	40588	0 mismatches/222 bases (0%)
Q9KX75	Putative protein, encoded on SCC elements. A reference sequence for this gene is from strain C427-ST42, GenBank ACSQ01000048.1 [310:813].	Reverse	6921	7423	40602	41105	0 mismatches/503 bases (0%)
Q7A207	Putative protein, encoded on SCC elements. A reference sequence for this gene is from strain C427-ST42, GenBank ACSQ01000048.1 [829:1143].	Reverse	7439	7750	41120	41432	0 mismatches/312 bases (0%)
Q7A206	Putative protein, encoded on SCC elements. A reference sequence for this gene is from strain C427-ST42, GenBank ACSQ01000048.1 [1230:1580].		7752	7838	41433	41520	0 mismatches/87 bases (0%)
ccrB-1	Cassette chromosome recombinase B, type 1. A reference sequence for this gene is from strain JCSC6690, GenBank AB705452.1 [7432:9057:r].	Reverse	8654	10282	42335	43964	0 mismatches/1629 bases (0%)
ccrA-1	Cassette chromosome recombinase A, type 1. 1268/1352 (94%) identities and 4/1352 gaps compared to strain JCSC6690 SCCmec type IX, GenBank AB705452.1 [10431-10962].	Reverse	10303	11652	43984	45334	0 mismatches/1350 bases (0%)
ORF No KK12	513/532 (96%) identities and 1/532 gaps compared to strain JCSC6690 SCCmec type IX, GenBank AB705452.1 [10431-10962].		11655	12185	45336	45867	0 mismatches/531 bases (0%)
cch	Cassette chromosome helicase.	Reverse	12188	13974	45868	47656	0 mismatches/1788 bases (0%)
orf7795	Putative protein from ACME element of strain C427. A reference sequence for this gene is from strain C427-ST42, GenBank ACSQ01000048.1 [7795:9354].	Reverse	14398	15957	48079	49639	0 mismatches/1560 bases (0%)
IR_IS431	Inverted repeat of IS431.		16224	16239	49905	49921	0 mismatches/16 bases (0%)
tnpIS431-06	Transposase for IS431.	Reverse	16284	16958	49965	50640	0 mismatches/675 bases (0%)
"DLJ55_14705"	Hypothetical protein. 97% identity compared to, GenBank CP029627.1:2809799-2816653, strain MOK042; 98% identity compared to CP029650.1:1401-8447, a plasmid from strain AR_0471, GenBank.	Forward	17243	24273	50924	57955	Cannot be assessed because of discrepancies between Nanopore and Illumina sequences.

Gene	Description/gene product and comments	Orientation	Start position in SCC	End position in SCC	Start position in genome	End position in genome	Comparison of ANR570100 to 02T-671
IR_IS431	Inverted repeat of IS431.		24603	24618	58284	58300	0 mismatches/16 bases (0%)
F8WKF9	Putative membrane protein. A reference sequence for this gene is from strain C427-ST42, GenBank ACSQ01000050 [548:900].	Forward	24821	25173	58502	58855	0 mismatches/353 bases (0%)
EHQ67276	Putative protein, branched-chain amino acid transport domain protein. A reference sequence for this gene is from strain C427-ST42, GenBank ACSQ01000050 [1158:1583].	Forward	25431	25856	59112	59538	0 mismatches/426 bases (0%)
***opp3A*/A8YYZ6**	Putative S-adenosyl-L-methionine-dependent methyltransferase from SCC elements. A reference sequence for this gene is from strain FPR3757, GenBank CP000255.1 [77130:77945].	Forward	26265	26903	59946	60585	0 mismatches/639 bases (0%)
opp3B	Nickel/peptide ABC superfamily ATP binding cassette transporter, membrane protein	Forward	28456	29412	62137	63094	0 mismatches/957 bases (0%)
opp3C	Oligopeptide permease, channel-forming protein	Forward	29412	30179	63093	63861	0 mismatches/768 bases (0%)
***opp3D*/A8YZ00**	Nickel/peptide ABC superfamily ATP binding cassette transporter, ABC protein, known from ACME elements.	Forward	30146	30913	63827	64595	0 mismatches/768 bases (0%)
***opp3E*/A8YZ01**	Nickel/peptide ABC superfamily ATP binding cassette transporter, ABC protein, known from ACME elements.	Forward	30907	31540	64587	65222	0 mismatches/635 bases (0%)
tnp_A8YYY6	Transposase.		31627	32293	65308	65975	0 mismatches/667 bases (0%)
DR_SCC	Direct repeat of SCC.		33683	33701	67364	67383	0 mismatches/19 bases (0%)

Table 3: Genes in Deviating Region 2 in comparison to canonical CC80, to 02T-671 and to canonical CC1.

	Comparison of ANR570100 to CC80 reference sequence 11819-97 (CP003194)	**Comparison of ANR570100 to CC1 reference sequence MW2 (BA000033)**	**Comparison of ANR570100 to 02T-671**	**Provenance in ANR570100**
salA	0 mismatches/1065 bases (0%)	0 mismatches/1065 bases (0%)	0 mismatches/1065 bases (0%)	CC1 or CC80
ycnB	1 mismatches/1443 bases (0.07%)	0 mismatches/1443 bases (0%)	0 mismatches/1443 bases (0%)	CC1 or CC80
sepA	1 mismatches/468 bases (0.21%)	1 mismatches/468 bases (0.21%)	0 mismatches/468 bases (0%)	CC1 or CC80
sdrM	4 mismatches/1344 bases (0.3%)	0 mismatches/1344 bases (0%)	0 mismatches/1344 bases (0%)	CC1 or CC80
hlIII	3 mismatches/684 bases (0.44%)	3 mismatches/684 bases (0.44%)	0 mismatches/684 bases (0%)	CC1 or CC80
urtF	3 mismatches/1188 bases (0.25%)	3 mismatches/1188 bases (0.25%)	0 mismatches/1188 bases (0%)	CC1 or CC80
yvsG	0 mismatches/516 bases (0%)	0 mismatches/516 bases (0%)	0 mismatches/516 bases (0%)	CC1 or CC80
ynzG	0 mismatches/261 bases (0%)	0 mismatches/261 bases (0%)	0 mismatches/261 bases (0%)	CC1 or CC80
Q5HE31	3 mismatches/1371 bases (0.22%)	0 mismatches/1371 bases (0%)	0 mismatches/1371 bases (0%)	CC1 or CC80
htsC	0 mismatches/969 bases (0%)	0 mismatches/969 bases (0%)	0 mismatches/969 bases (0%)	CC1 or CC80
htsB	10 mismatches/1032 bases (0.97%)	0 mismatches/1032 bases (0%)	0 mismatches/1032 bases (0%)	CC1
htsA	0 mismatches/984 bases (0%)	0 mismatches/984 bases (0%)	0 mismatches/984 bases (0%)	CC1 or CC80
tnpIS1	*Absent from CP003194*	2 deletions in MW2/34 bases (5.88%)	0 mismatches/34 bases (0%)	CC1
Q5HE27	4 mismatches/1071 bases (0.37%)	0 mismatches/1071 bases (0%)	0 mismatches/1071 bases (0%)	CC1
rhbC1	18 mismatches/1758 bases (1.02%)	0 mismatches/1758 bases (0%)	1 mismatches/1757 bases (0.06%)	CC1
Q5HE25	13 mismatches/1194 bases (1.09%)	2 mismatches/1194 bases (0.17%)	1 mismatches/1194 bases (0.08%)	CC1
rhbC2	16 mismatches/1977 bases (0.81%)	0 mismatches/1977 bases (0%)	0 mismatches/1977 bases (0%)	CC1
asp23	12 mismatches/510 bases (2.35%)	1 mismatches/510 bases (0.2%)	0 mismatches/510 bases (0%)	CC1
DUF2273	0 mismatches/240 bases (0%)	0 mismatches/240 bases (0%)	0 mismatches/240 bases (0%)	CC1 or CC80
Q5HE21	1 mismatches/549 bases (0.18%)	2 mismatches/549 bases (0.36%)	0 mismatches/549 bases (0%)	CC1 or CC80
opuD2	20 mismatches/1563 bases (1.28%)	0 mismatches/1563 bases (0%)	0 mismatches/1563 bases (0%)	CC1
Q5HE19	0 mismatches/1008 bases (0%)	0 mismatches/1008 bases (0%)	0 mismatches/1008 bases (0%)	CC1 or CC80
qorA	6 mismatches/1002 bases (0.6%)	0 mismatches/1002 bases (0%)	0 mismatches/1002 bases (0%)	CC1
DUF915	1 mismatches/870 bases (0.11%)	1 mismatches/870 bases (0.11%)	0 mismatches/870 bases (0%)	CC1 or CC80
lacG	2 mismatches/1413 bases (0.14%)	0 mismatches/1413 bases (0%)	0 mismatches/1413 bases (0%)	CC1
lacE	0 mismatches/1719 bases (0%)	10 mismatches/1719 bases (0.58%)	0 mismatches/1719 bases (0%)	CC1 or CC80
lacF	6 mismatches/312 bases (1.92%)	0 mismatches/312 bases (0%)	0 mismatches/312 bases (0%)	CC1
lacD	20 mismatches/981 bases (2.04%)	1 mismatches/981 bases (0.1%)	0 mismatches/981 bases (0%)	CC1
lacC	4 mismatches/933 bases (0.43%)	4 mismatches/933 bases (0.43%)	0 mismatches/933 bases (0%)	CC1 or CC80
lacB	4 mismatches/516 bases (0.78%)	3 mismatches/516 bases (0.58%)	0 mismatches/516 bases (0%)	CC1
lacA	6 mismatches/429 bases (1.4%)	0 mismatches/429 bases (0%)	0 mismatches/429 bases (0%)	CC1
tx_lacR	5 mismatches/68 bases (7.35%)	0 mismatches/68 bases (0%)	0 mismatches/68 bases (0%)	CC1
lacR	6 mismatches/756 bases (0.79%)	0 mismatches/756 bases (0%)	0 mismatches/756 bases (0%)	CC1
cobB	5 mismatches/732 bases (0.68%)	5 mismatches/732 bases (0.68%)	0 mismatches/732 bases (0%)	CC1 or CC80

	Comparison of ANR570100 to CC80 reference sequence 11819-97 (CP003194)	Comparison of ANR570100 to CC1 reference sequence MW2 (BA000033)	Comparison of ANR570100 to 02T-671	Provenance in ANR570100
D9RC03	*Absent from CP003194 (100%)*	0 mismatches/98 bases (0%)	0 mismatches/98 bases (0%)	CC1
DUF3885	*Absent from CP003194 (100%)*	14 mismatches/609 bases (2.3%)	14 mismatches/609 bases (2.3%)	Unknown, but other than CC80
D9RC05	*Absent from CP003194 (100%)*	0 mismatches/249 bases (0%)	0 mismatches/249 bases (0%)	CC1
Q5HE05	*Absent from CP003194 (100%)*	0 mismatches/126 bases (0%)	0 mismatches/126 bases (0%)	CC1
yvgN2	6 mismatches/849 bases (0.71%)	0 mismatches/849 bases (0%)	0 mismatches/849 bases (0%)	CC1
adhR	2 mismatches/417 bases (0.48%)	0 mismatches/417 bases (0%)	0 mismatches/417 bases (0%)	CC1
hysA	201 mismatches/2434 bases (8.19%)	0 mismatches/2448 bases (0%)	1 mismatches/2448 bases (0.04%)	CC1
att_nyEtd	*Absent from ANR570100*	*Absent from ANR570100 and MW2*	*Absent from ANR570100 and 02T-671*	Absent (as in CC1, unlike CC80)
hsdS-etd	*Absent from ANR570100*	*Absent from ANR570100 and MW2*	*Absent from ANR570100 and 02T-671*	Absent (as in CC1, unlike CC80)
hsdM	*Absent from ANR570100*	*Absent from ANR570100 and MW2*	*Absent from ANR570100 and 02T-671*	Absent (as in CC1, unlike CC80)
etD	*Absent from ANR570100*	*Absent from ANR570100 and MW2*	*Absent from ANR570100 and 02T-671*	Absent (as in CC1, unlike CC80)
F3TKB7-var1	*Absent from ANR570100*	*Absent from ANR570100 and MW2*	*Absent from ANR570100 and 02T-671*	Absent (as in CC1, unlike CC80)
edinB	*Absent from ANR570100*	*Absent from ANR570100 and MW2*	*Absent from ANR570100 and 02T-671*	Absent (as in CC1, unlike CC80)
F5W4X2	*Absent from ANR570100*	*Absent from ANR570100 and MW2*	*Absent from ANR570100 and 02T-671*	Absent (as in CC1, unlike CC80)
att_nyEtd	*Absent from ANR570100*	*Absent from ANR570100 and MW2*	*Absent from ANR570100 and 02T-671*	Absent (as in CC1, unlike CC80)
Q5HE00	*Absent from ANR570100*	*Absent from ANR570100 and MW2*	*Absent from ANR570100 and 02T-671*	Absent (as in CC1, unlike CC80)
eapH-1	6 mismatches/426 bases (1.41%)	0 mismatches/426 bases (0%)	0 mismatches/426 bases (0%)	CC1
alsD-L1	12 mismatches/705 bases (1.7%)	0 mismatches/705 bases (0%)	0 mismatches/705 bases (0%)	CC1
alsS	32 mismatches/1665 bases (1.92%)	0 mismatches/1665 bases (0%)	0 mismatches/1665 bases (0%)	CC1
Q8NVB9	4 mismatches/183 bases (2.19%)	0 mismatches/183 bases (0%)	0 mismatches/183 bases (0%)	CC1
rpsI	0 mismatches/399 bases (0%)	0 mismatches/399 bases (0%)	0 mismatches/399 bases (0%)	CC1 or CC80
rplM	1 mismatches/438 bases (0.23%)	1 mismatches/438 bases (0.23%)	1 mismatches/438 bases (0.23%)	CC1 or CC80
L13_leader	0 mismatches/70 bases (0%)	0 mismatches/70 bases (0%)	0 mismatches/70 bases (0%)	CC1 or CC80
truA	14 mismatches/804 bases (1.74%)	0 mismatches/804 bases (0%)	0 mismatches/804 bases (0%)	CC1
ecfT	6 mismatches/807 bases (0.74%)	0 mismatches/807 bases (0%)	0 mismatches/807 bases (0%)	CC1
ecfA1	6 mismatches/861 bases (0.7%)	0 mismatches/861 bases (0%)	0 mismatches/861 bases (0%)	CC1
ecfA2	8 mismatches/810 bases (0.99%)	0 mismatches/810 bases (0%)	0 mismatches/810 bases (0%)	CC1
rplQ	0 mismatches/369 bases (0%)	0 mismatches/369 bases (0%)	0 mismatches/369 bases (0%)	CC1 or CC80
rpoA	0 mismatches/945 bases (0%)	0 mismatches/945 bases (0%)	0 mismatches/945 bases (0%)	CC1 or CC80
rpsK	1 mismatches/390 bases (0.26%)	0 mismatches/390 bases (0%)	0 mismatches/390 bases (0%)	CC1 or CC80
rpsM	0 mismatches/366 bases (0%)	0 mismatches/366 bases (0%)	0 mismatches/366 bases (0%)	CC1 or CC80
rpmJ	0 mismatches/114 bases (0%)	0 mismatches/114 bases (0%)	0 mismatches/114 bases (0%)	CC1 or CC80
infA	0 mismatches/219 bases (0%)	0 mismatches/219 bases (0%)	0 mismatches/219 bases (0%)	CC1 or CC80
adk	0 mismatches/648 bases (0%)	2 mismatches/648 bases (0.31%)	0 mismatches/648 bases (0%)	CC1 or CC80
secY1	1 mismatches/1293 bases (0.08%)	1 mismatches/1293 bases (0.08%)	0 mismatches/1293 bases (0%)	CC1 or CC80
rplO	1 mismatches/441 bases (0.23%)	0 mismatches/441 bases (0%)	0 mismatches/441 bases (0%)	CC1 or CC80
rpmD	0 mismatches/180 bases (0%)	0 mismatches/180 bases (0%)	0 mismatches/180 bases (0%)	CC1 or CC80
rpsE	0 mismatches/501 bases (0%)	0 mismatches/501 bases (0%)	0 mismatches/501 bases (0%)	CC1 or CC80
rplR	0 mismatches/360 bases (0%)	0 mismatches/360 bases (0%)	0 mismatches/360 bases (0%)	CC1 or CC80

	Comparison of ANR570100 to CC80 reference sequence 11819-97 (CP003194)	Comparison of ANR570100 to CC1 reference sequence MW2 (BA000033)	Comparison of ANR570100 to 02T-671	Provenance in ANR570100
rplF	0 mismatches/537 bases (0%)	0 mismatches/537 bases (0%)	0 mismatches/537 bases (0%)	CC1 or CC80
rpsH	0 mismatches/399 bases (0%)	1 mismatches/399 bases (0.25%)	0 mismatches/399 bases (0%)	CC1 or CC80
rpsZ	0 mismatches/186 bases (0%)	0 mismatches/186 bases (0%)	0 mismatches/186 bases (0%)	CC1 or CC80
rplE	0 mismatches/540 bases (0%)	0 mismatches/540 bases (0%)	1 mismatches/540 bases (0.19%)	CC1 or CC80
rplX	0 mismatches/318 bases (0%)	0 mismatches/318 bases (0%)	0 mismatches/318 bases (0%)	CC1 or CC80
rplN	0 mismatches/369 bases (0%)	0 mismatches/369 bases (0%)	0 mismatches/369 bases (0%)	CC1 or CC80
rpsQ	0 mismatches/264 bases (0%)	0 mismatches/264 bases (0%)	0 mismatches/264 bases (0%)	CC1 or CC80
rpmC	0 mismatches/210 bases (0%)	0 mismatches/210 bases (0%)	0 mismatches/210 bases (0%)	CC1 or CC80

Figure 1: Schematic diagram of the genomes of 02T567 (outer circle), ANR570100 (middle circle) and the reference genome CC80, 11819-97, GenBank CP003194.1 (inner circle). Genomic fragments are colour-coded depending on their origin.