

Title: Comparative genomics suggests a taxonomic revision of the *Staphylococcus cohnii* species complex

Anna Lavecchia^{1†}, Matteo Chiara^{1,2†}, Caterina De Virgilio³, Caterina Manzari¹, Carlo Pazzani⁴, David Horner^{1,2}, Graziano Pesole^{1,3,5} and Antonio Placido^{1*}

¹ Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies, National Research Council, Bari, Italy.

² Department of Biosciences, University of Milan, Milan, Italy;

³ Department of Biosciences, Biotechnology and Biopharmaceutics, University of Bari “Aldo Moro”, Bari, Italy;

⁴ Department of Biology, University of Bari Aldo Moro, Bari, Italy.

⁵ Consorzio Interuniversitario Biotecnologie, Trieste, Italy.

†These authors contributed equally to this work.

*Corresponding author: E-mail: a.placido@ibiom.cnr.it

Abstract: *Staphylococcus cohnii* (SC), a coagulase-negative bacterium, was first isolated in 1975 from human skin. Early phenotypic analyses led to the delineation of two subspecies (subsp.), *Staphylococcus cohnii* subsp. *cohnii* (SCC) and *Staphylococcus cohnii* subsp. *urealyticus* (SCU). SCC was considered to be specific to humans whereas SCU apparently demonstrated a wider host range, from lower primates to humans. The type strains ATCC 29974 and ATCC 49330 have been designated for SCC and SCU, respectively. Comparative analysis of 66 complete genome sequences – including a novel SC isolate – revealed unexpected patterns within the SC complex, both in terms of genomic sequence identity and gene content, highlighting the presence of 3 phylogenetically distinct groups. Based on our observations, and on the current guidelines for taxonomic classification for bacterial species, we propose a revision of the SC species complex. We suggest that SCC and SCU should be regarded as two distinct species: SC and SU (*Staphylococcus urealyticus*), and that two distinct subspecies, SCC and SCB (SC subsp. *barensis*, represented by the novel strain isolated in Bari) should be recognized within SC. Furthermore, since large scale comparative genomics studies recurrently suggest inconsistencies or conflicts in taxonomic

assignments of bacterial species, we believe that the approach proposed here might be considered for more general application.

Keywords: Comparative genomics; Genome Shotgun Sequencing; DNA-DNA hybridization analyses; Average Nucleotide Identity; Phylogenetic analyses; *Staphylococcus cohnii*.

Significance statement

In recent years, as the extent of its involvement in a multitude of human and animal infections has become evident, much research has been focused on *Staphylococcus cohnii*. Moreover, *S. cohnii* is also widely used as a model in the development of biotechnological applications and antibacterial medical devices. Its relevance notwithstanding, comparative genomic studies of this species complex are lacking, and the current work suggests the need for a major taxonomic revision.

More generally, the computational approach presented here can be applied on a larger scale, both for the resolution of complex or conflicting taxonomic assignments and as a tool to contribute to the understanding of the history of biomedically and biotechnologically important traits at species and subspecies levels.

Introduction

Staphylococcus cohnii (SC), a Gram-positive bacterium of the Coagulase-Negative Staphylococci (CoNS) group, was first isolated by Schleifer and Kloos in 1975. The name *conhii* was adopted in memory of Ferdinand Julius Cohn, a German botanist and bacteriologist (Schleifer & Kloos 1975).

According to the current classification - which is fundamentally based on phenotypic traits - SC includes two subspecies (subsp.): *Staphylococcus cohnii* subsp. *cohnii* (SCC) and *Staphylococcus cohnii* subsp. *urealyticum* (SCU). The SCC ATCC 29974 and SCU ATCC 49330 isolates have been designated as the type strains for SCC and SCU respectively (Kloos & Wolfshohl 1983, 1991). The original spelling, SC subsp. *urealytium* (sic), was corrected by Sneath to SC subsp. *urealyticus* (SCU) in 1992 (Sneath 1992).

A larger colony size, distinct pigmentation, differences in fatty acid profile and the presence of metabolic activities including β -glucuronidase and β -galactosidase activities, delayed alkaline phosphatase activity, and the ability to produce acid aerobically from α -lactose discriminate SCU from SCC. Moreover, SCC was originally reported to colonize only humans while SCU can also colonize other primates (Kloos & Wolfshohl 1991). More recently SCU has also been isolated from healthy dogs (Bean et al. 2017) and goats (Seni et al. 2019).

Similar to most CoNS, SCC and SCU are typically commensal bacteria of the skin and mucous membranes (Waldon et al. 2002; Crossley et al. 2009). However, several opportunistically pathogenic strains have been described and implicated in nosocomial infections, including meningitis, primary septic arthritis, septicaemia, brain abscess, and catheter invasion (Okudera et al. 1991; Mastroianni et al. 1996; Basaglia et al. 2003; Adeyemi et al. 2010; Yamashita et al. 2005; Mendoza-Olazarán et al. 2017). The ability to form biofilms appears to play an important role in staphylococcal virulence (Yong et al. 2019), and biofilm-associated infections are of particular concern because they are often difficult to resolve with antibiotics. Recent studies, suggest that similar to other CoNS, SC can adhere to and invade human HeLa cells, through the formation of biofilms (Szczuka et al. 2016), while strains of SC and especially those isolated from hospital environments, including paediatric wards and intensive-care units, have been reported to be resistant to several antibiotics (Szewczyk et al. 2000, 2004; Song et al. 2017). Indeed, multidrug-resistant CoNS bacteria, constitute an emerging source of concern for Public Health organizations (Moawad et al. 2019; David & Elliott 2015) as they have been associated with an increasing proportion of nosocomial infections, and because they can act as a reservoir of resistance determinants for *Staphylococcus aureus* through horizontal gene transfer (Otto 2013; Winstel et al. 2013; Larsen et al. 2017; Argemi et al. 2019).

Comparative genomic and phylogenetic studies allow the characterization of evolutionary dynamics and the identification of genes and pathways potentially involved in pathogenesis and/or antibiotic resistance. Here we present comprehensive analyses of the complete collection of 65 publicly available SC genomes, as well as that of a novel strain. We uncover striking patterns of genomic evolution, including high levels of genomic diversity and differential gene acquisition and loss that suggest a taxonomic revision of the SC species complex. We propose that SC should be divided into two species, SC and SU (*Staphylococcus urealyticus*). Moreover, two subspecies SCC and SCB (SC subsp. *barensis*, exemplified by our novel strain isolated in Bari) should be distinguished within SC. Of more general interest, we describe an approach based on the integration of different types of phylogenetic, genomic and gene content analyses that could be applied on a larger scale for the resolution of complex or conflicting taxonomic assignments.

Material and Methods

Isolation of SC5 and preliminary taxonomic detection

A Nunc bioassay plate containing LB (Luria Bertani) agar supplemented with 0.2 mM Potassium Chromate and Chloramphenicol (12.5 µg/ml) was prepared and placed on the worktop of a class II biological safety cabinet.

Airflow was switched on for one hour, and the plate was subsequently incubated for 16 hours at 37°C. Sixty-two colonies were grown.

All colonies were replicated on LB agar containing Cr⁺⁶ 150 mM. Only five colonies survived (Lavecchia et al. 2018), and were subjected to a preliminary taxonomic characterization through partial 16S rDNA amplification and Sanger sequencing. One colony preliminarily assigned to *Staphylococcus cohnii* and labeled SC5 was then subjected to whole genome sequencing.

PCR amplification of the SC5 16S rDNA was performed using 200 ng of DNA (see DNA isolation section) and the following primers: 1 µM Forward 5'-TACGGGAGGCAGCAGTAG-3' (16S rDNA position 369-386), 1 µM Reverse 5'-CATGGTGTGACGGGCG GT-3' (position 1424-1441). The reaction mixture (final volume 50 µl) was completed using: each NTP 200 µM, 2 mM MgCl₂, and 2 U Taq DNA polymerase (Thermo Fisher Scientific, Massachusetts, USA). Taq DNA polymerase was activated as follows: 94 °C for 5 min; 30 cycles at 94 °C for 30 sec, 55 °C for 15 sec, and 72 °C for 1 min.

Sanger sequencing was performed by Macrogen (Amsterdam, Netherlands) and the preliminary taxonomic assessment was made by probing the 16S rRNA (Bacteria and Archaea) database available at the NCBI (<https://blast.ncbi.nlm.nih.gov/Blast>), using the BLASTn algorithm.

DNA isolation, library preparation and sequencing

Genomic DNA was extracted with the DNeasy Blood and Tissue kit (Qiagen, Hilden Germany). The library was prepared using the Nextera XT library prep workflow (Illumina, Illumina San Diego, CA) and 2 x 250 nt paired-end reads generated on an Illumina MiSeq instrument.

Genome assembly and annotation

Raw data were processed using a modified version of the “Fosmid1” pipeline in the A-GAME (Chiara et al. 2018) Galaxy framework (Afgan et al. 2018). Quality trimming was executed using the sliding-window operation in Trimmomatic with default parameters (Bolger et al. 2014). Overlapping reads were merged using PEAR with standard parameters (Zhang et al. 2014). The final assembly was performed using the SPAdes assembler (version 3.50) using kmers of 33, 55, 77, 99, and 121 nt (Bankevich et al. 2012). Annotation was performed with PROKKA using default parameters (Seemann 2014).

SC genomes used in this study and annotation of protein coding genes

The complete collection of 65 SC genomes assemblies (including *Staphylococcus cohnii* subsp. *cohnii* ATCC 29974 and *Staphylococcus cohnii* subsp. *urealyticus* ATCC 49330), as available in Genbank on July 1st 2020, was downloaded from the NCBI assembly database, directly from the “Download Assemblies” link, as available from the web interface. To avoid possible ascertainment biases, all the genomes were re-annotated using the procedure described above. Annotations of protein coding genes, as obtained from Prokka, were used in all the subsequent analyses. A complete list of the accession number of the genomes used in this study is provided in supplementary Table S1.

Calculation of Average Nucleotide Identity (ANI) and *in silico* DDH (DNA-DNA hybridization)

ANI based on BLAST (ANiB) between all 66 genomes (SC5 included) was computed according to the method described by Rossello-Moras (Richter & Rosselló-Móra 2009), as implemented by a custom script, which is available at https://github.com/cvulpispaper/compute_anib. *In silico* DDH was computed using the GGDC (Genome-to-Genome Distance Calculator 2.1) available from <https://ggdc.dsmz.de/ggdc.php#> (Meier-Kolthoff et al. 2013). As recommended in (Auch et al, 2010), all the comparisons performed in this study were based on formula 2.

Clusters of orthologous genes, core and accessory genome

The makeblastdb utility, as incorporated in the blast+ software package (Camacho et al, 2008) was used to prepare a blast protein database, containing all of the protein coding genes as predicted by Prokka in the 66 SC genomes included in this study. All-against-all BLASTp (Altschul et al. 1990) was performed using the BLOSUM80 matrix and accepting only best reciprocal hits with e-value $\leq 1e-5$ and where “second-best” hits from the same genome produce bit scores less than 90% of that associated with the best match. Putative clusters of orthologous genes (COGs) were established as groups of best reciprocal BLAST hits. Core genes were defined as COGs containing single representatives from all genomes included in our analyses (or all genomes within major groups), and accessory genes as COGs with incomplete representation. The program used for the identification of COGs is available at https://github.com/cvulpispaper/compute_aai_and_cogs.

Estimation of completeness of core and accessory genomes

Size of core and accessory genomes were established by rarefaction analyses based on random resampling of genomic sequences of each major group. For each number of strains considered (2–28 for B, 2 to 22 for A1, 2 to 16 for A2, and 2 to 66 for SC) the inferred sizes of core and accessory genomes were recorded for 10,000 replicates of randomly selected combinations of genomes. Plots were prepared showing mean and standard deviation of these statistics.

Phyletic patterns and clustering of gene absence presence profiles

The phyletic pattern of genes presence/absence in the genomes of the 66 SC isolates was inferred directly by comparison of clusters of orthologous genes. Only COGs containing 10 or more genes were considered in this analysis. A matrix of gene absence/presence was compiled, with genes on the rows and isolates on the columns. A value of 0 was used to indicate the absence of a gene, a value of 1 its presence. A correlation-based distance matrix of gene absence/presence profiles was obtained by applying the `cor` and the `dist` functions, from the `stat` library of the R programming language with default parameters (Pearson correlation and Euclidean distances respectively). Clustering was performed by applying the `hclust` function with median linkage, from the same software package.

Phylogenetic analyses

The conceptually translated sequences of the 1468 SC core genes were independently aligned using Muscle (Edgar 2004) and ambiguously aligned regions were excluded using the GBlocks software (Castresana 2000). Maximum-likelihood phylogenetic reconstruction and bootstrap analyses of concatenated alignments were performed using the software PHYML (Guindon et al. 2009) under the WAG (Whelan & Goldman 2001) substitution model, suggested by the software ProtTest (Darriba et al. 2011) to best fit the data, with invariable and four gamma-distributed substitution rate categories.

Statistical analyses

Welch T-test p-values for the comparison of ANIb distributions and the size of the core and accessory genomes were computed by means of the `t.test` function as implemented in the `stats` R package.

Results

Isolation and Whole-Genome Shotgun Sequencing of a novel strain of SC (*Staphylococcus cohnii*)

Five strains of staphylococci were isolated from a disused class II biological safety cabinet during a study aiming to identify bacterial strains resistant to hexavalent chromate (Lavecchia et al. 2018). Preliminary taxonomic analyses based on partial 16S rDNA Sanger sequencing identified four of these strains as *Staphylococcus arlettae* (Lavecchia et al. 2018), while one isolate showed high levels of similarity (99.1 % and 98.8%) respectively with the type strains of SCC (*Staphylococcus cohnii* subsp. *cohnii*) and SCU (*Staphylococcus cohnii* subsp. *urealyticus*). The latter, preliminarily named SC5, was subsequently subjected to Whole-Genome Shotgun Sequencing using an Illumina MiSeq instrument. A total of 2,402,324 paired end reads were obtained, with an average insert size of 254.74 bp providing a theoretical 230x coverage of the genome. Raw reads were subjected to quality trimming and assembly by means of a modified version of the Fosmid1 pipeline as incorporated in A-GAME (Chiara et al. 2018). Salient features of the SC5 genome assembly are summarized in Table 1. An overall good level of contiguity was observed with more than 90% of the assembly incorporated in contigs greater than 100 Kb in size (N90 108 Kb). A total of 10 rRNA, 61 tRNA and 2510 protein coding genes were predicted by *in silico* annotation of the genome. Of note, the *emrA* and *emrB* genes, implicated in chromate and ampicillin co-resistance in *Staphylococcus aureus* LZ 01 (Zhang et al. 2016), were identified in the genome of SC5. These genes are also observed in the genomes of the other 4 *S. arlettae* chromium resistant strains isolated from the same environment. The draft genome sequence of SC5 was deposited in NCBI under the accession number JAALCY000000000, BioSample accession number SAMN14142771 and BioProject number ID PRJNA607668.

Table 1.

Main genome assembly features of SC5, SCC ATCC 29974 and SCU ATCC 49330 strains. DDH and ANIb values between strains.

Strain	Size (Mb)	GC (%)	Contigs	N50 (Kb)	Proteins	rRNAs	tRNAs	DDH (%) ^a			ANIb (%) ^b		
								SC5	SCC	SCU	SC5	SCC	SCU
SC5	2.62	32.2	34	806	2510	10	61	-	67.5	41.9	-	95.4	91.0
SCC ATCC 29974	2.71	32.6	83	114	2422	9	58	67.5	-	-	95.4	-	91.5
SCU ATCC 49330	2.67	32.5	223	26	2457	13	61	41.9	-	-	91.0	91.5	-

^a DDH, DNA-DNA hybridization (cut-off for species affiliation > 70%); ^b ANIb, Average Nucleotide Identity based on BLAST (cut-off for species affiliation > 96%). SCC ATCC29974, *Staphylococcus cohnii* subsp. *cohnii* type strain; SCU ATCC49330, *Staphylococcus cohnii* subsp. *urealyticus* type strain.

In silico DNA-DNA hybridization analyses

In silico DNA-DNA hybridization (DDH) analyses were performed to refine the taxonomic delineation of SC5. Strikingly, while 16S rDNA taxonomic assignment suggested that SC5 was closely related to SCC, *in silico* hybridization assays against the SCC ATCC29974 and SCU ATCC49330 type strains recovered somewhat unexpected patterns. Indeed (table 1), the observed DDH values, 67.5% and 41.9% respectively for SCC and SCU, were borderline or well below the cut-value of 70%, that is normally used to delineate species by this method (Meier-Kolthoff et al. 2013; Colston et al. 2014; Garrido-Sanz et al. 2016). A systematic comparison of *in silico* hybridization profiles of SC5 against the complete collection of the other 65 SC draft genomes considered in this study (supplementary table S1) was performed. Notably, contrasting patterns of sequence similarity profiles (supplementary table S2) were observed. SC5 showed DDH levels greater than 90% with 15 isolates (supplementary Figure S1), but lower than 70% with 22 (including the SCC ATCC29974 type strain) and around 45% with the remaining 28 isolates (including the SCU ATCC49330 type strain).

Analysis of genomic identity

Levels of pairwise genome identity between all currently available SC genomes were established by ANIb (Average Nucleotide Identity on BLAST). Hierarchical clustering of ANIb profiles, was applied to identify clades/groups of SC genomes with similar levels of genome identity. As shown in Figure 1 and supplementary Figure S2, and consistent with patterns of *in silico* DDH profiles, the results of these analyses suggested the presence of 3 distinct clusters with different mutual levels of ANIb within the SC species complex. The first two groups of isolates, referred to hereafter as A1 and A2, correspond to isolates SCC and are more closely related, while the third group (referred as B), is more distantly related to A1 and A2 and is composed exclusively of SCU isolates. A1 incorporates 22 strains, including the SCC type strain ATCC 29974, A2 is composed of 16 strains and includes SC5. Finally, group B contains 28 isolates, including the type strain SCU ATCC 49330.

Comparisons of genomic identity levels between the draft genomes of SCC and SCU show an average ANIb of 89.43%, a value that is well below the cut-off normally considered for inclusion in the same bacterial species (Varghese et al. 2015; Otto 2008). This approach also sustains the presence of two distinct clusters within SCC, with significantly different levels of ANIb (T-test p-value $\leq 1e-16$) and an average genome sequence identity of 95.82% (supplementary Figure S2), a value that is normally considered borderline for the identification of bacterial species (Varghese et al. 2015; Otto 2008). Taken together, our analyses of genomic similarity profiles by means of two independent methods strongly suggest that, according to the current guidelines for the delineation of bacterial species SCC and SCU should be considered as two distinct species. Consistent with this consideration

we observed that the ANI_b value recovered from the comparison of the 2 type strains SCC ATCC29974 and SCU ATCC49330 is 91.5%. Notably, comparisons between the draft genome assembly of SC5 with the SCC ATCC29974 and SCU ATCC49330 type strains resulted in ANI_b values of 95.4% and 91.0% respectively (table 1).

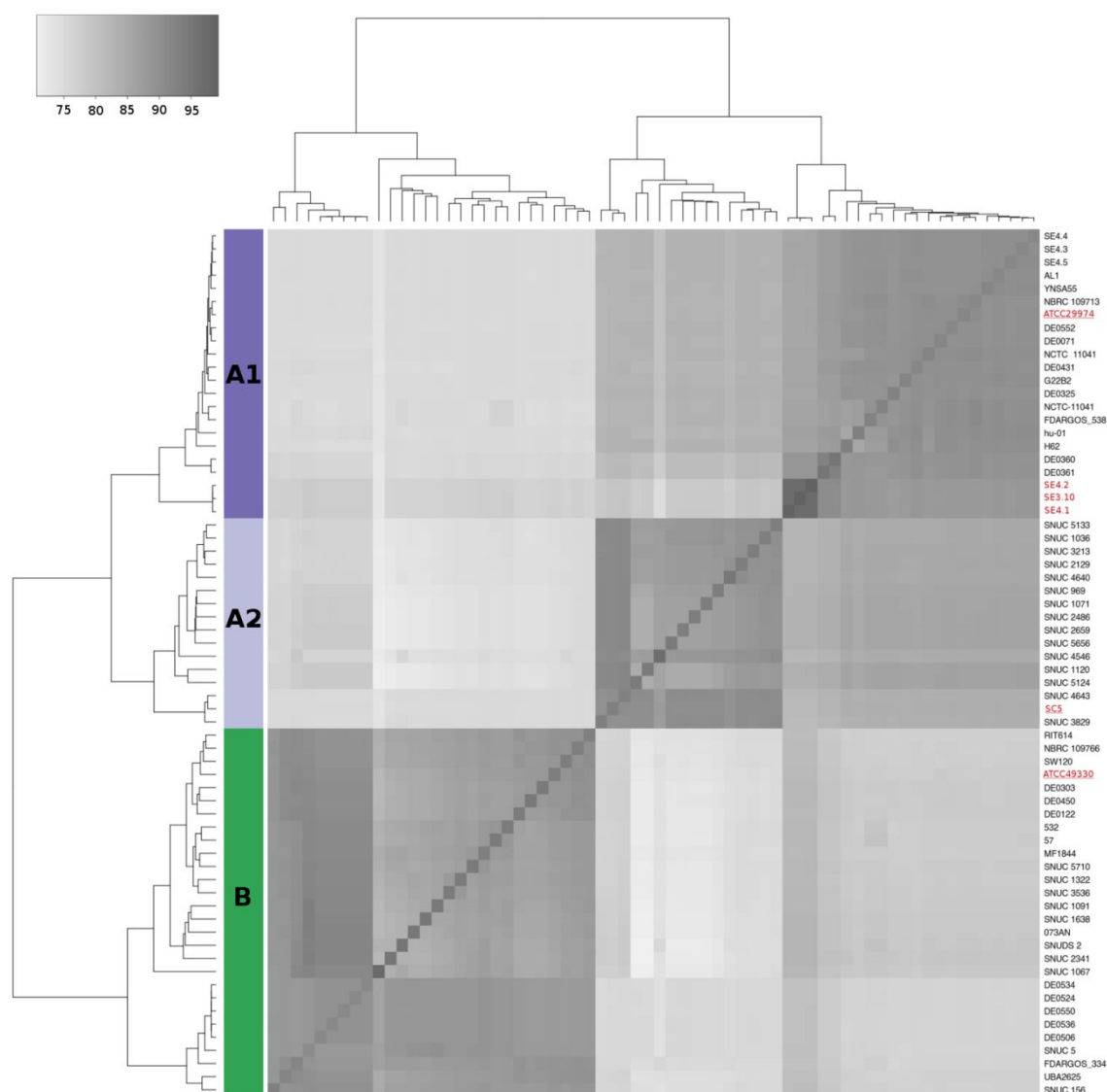


Fig. 1

Cluster of orthologous genes and phylogenetic analyses

Clusters of orthologous genes (COGs) as well as core and accessory genomes were established using an approach based on best reciprocal BLAST hits. A total of 5456 clusters of putative orthologs with more than one gene and 5044 singleton genes were identified. Hierarchical clustering of phenetic patterns of gene absence/presence

profiles was applied to identify SC isolates with a similar gene content. Consistent with our previous observations, three distinct groups were observed, identical in size and composition to A1, A2 and B. Notably, while the A1 and A2 clades were delineated very clearly by this analysis (fig. 2), suggesting a similar gene content within isolates of these groups, a somewhat more heterogeneous pattern was observed for group B suggesting a higher plasticity of the pan-genome, possibly associated with lateral gene transfer.

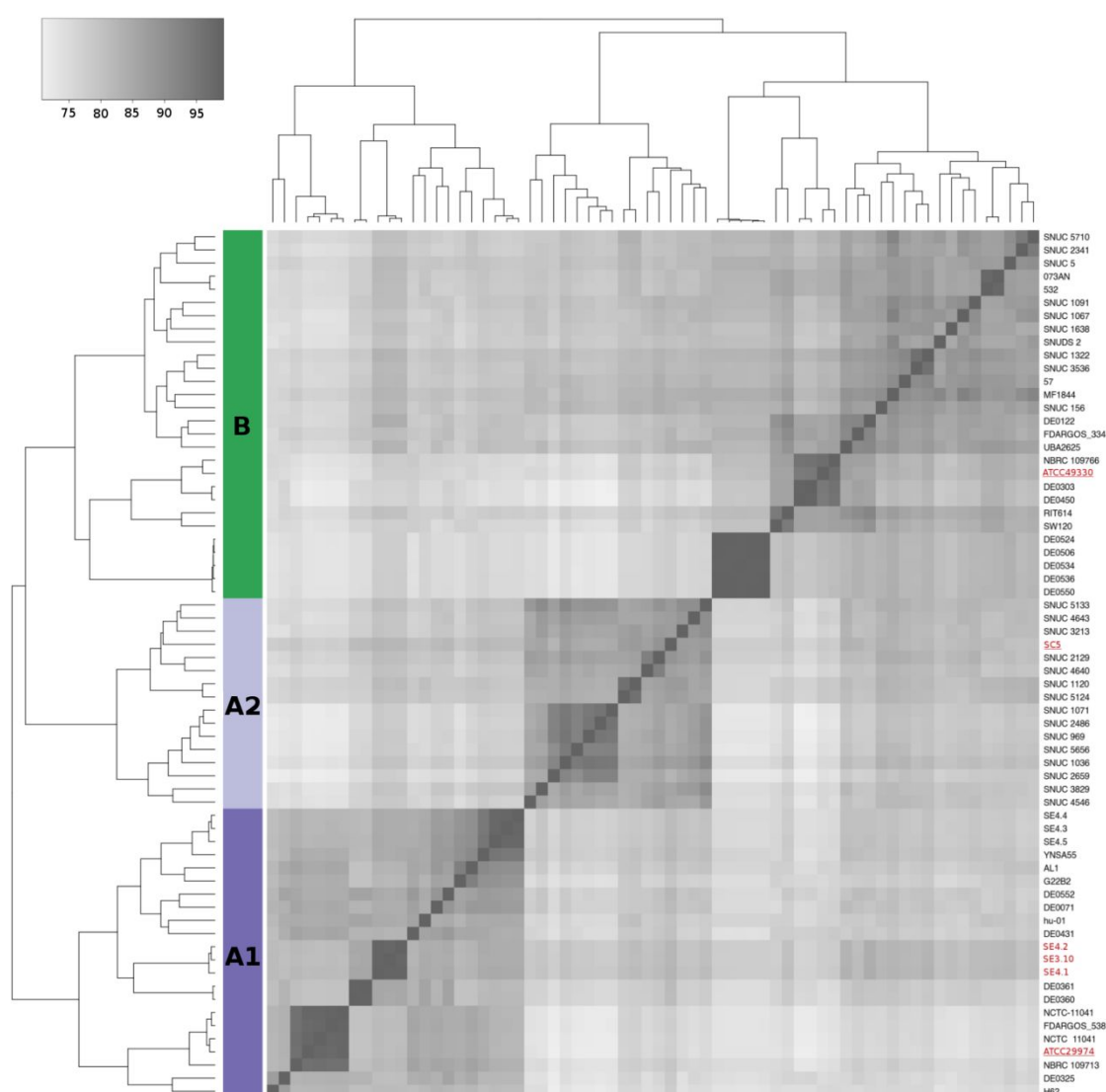


Fig. 2

Analysis of the SC core genome provides additional evidence for the different gene content in the A1, A2 and B. Indeed, when all the 66 available genomes are considered, a core genome of 1468 genes was recovered, while notable differences were observed in the size of the core genome between the 3 groups (fig. 3A).

While the estimated core genome size for the A2 cluster was 1889 genes, the core genomes of A1 and B were larger, with an estimated sizes of 1993 and 1953 genes respectively. Since our analyses are based on nearly equivalent numbers of isolates for every group, this difference is unlikely to be the result of a biased sampling, but might reflect a tendency for a more compact genome with a reduced number of genes in the A2 clade. Consistent with this hypothesis we observe that the average number of predicted protein coding genes is significantly reduced in A2 (average 2566) with respect to A1 (2618) and B (2651), with p-values of 0.06 and 0.022 respectively according to a Welch T-Test.

While our analyses suggest that the accessory genome of SC is relatively open (fig. 3B) and that additional genes are likely to be discovered as new genomic sequences become available, we note once again that the accessory genome in A2 is substantially reduced with respect to A1 and B, again consistent with systematic differences in gene content.

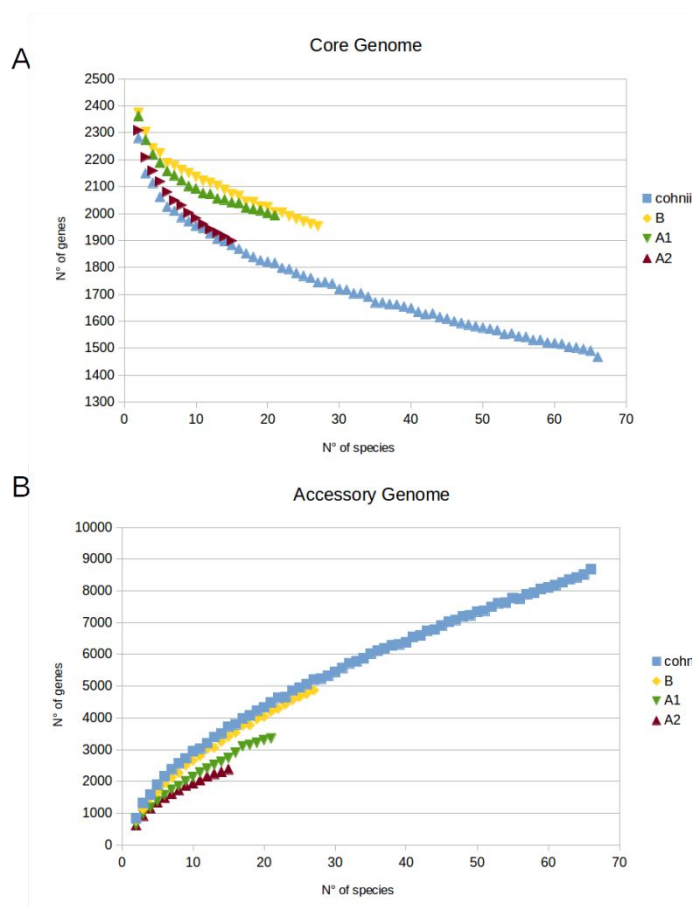


Fig. 3

Phylogenetic analyses of the concatenated alignment of 1468 core genes (fig. 4), recovered a tree with a topology consistent with the clustering of the isolates based on genome identity levels, providing an additional line of evidence for the presence of 3 distinct clades within the SC species complex. However, we notice that according

to our phylogenetic analyses, A1 is not strictly monophyletic and a distinct basal clade formed by 3 isolates: SE4.1, SE 4.2 and SE3.10 is observed. Interestingly, a similar pattern is replicated also in Figure 1, where the same group of isolates (SE4.1, SE 4.2 and SE3.10) form an identical basal clade, suggesting overall reduced levels of genome identity of these three strains with the other strains included in A1. Notably, the same pattern is not recovered when clustering of isolates based on gene content is considered (Figure 2). This indicates that the gene content of SE4.1, SE 4.2 and SE3.10 is highly consistent with that of other strains in the A1 cluster. Taken all together these observations might suggest widespread lateral gene transfer between SE4.1, SE 4.2 and SE3.10 and other SC isolates or alternatively, that while having a similar gene content with strains included in the A1 group, overall SE4.1, SE 4.2 and SE3.10 are highly divergent at sequence level. Possibly indicating faster evolutionary rates and/or ongoing diversifying selection. Importantly, we underscore that these 3 strains were isolated from a similar environment (rice seeds), and geographic location (India), consistent with the hypothesis that the observed reduction in genome/protein identity levels might reflect the regional/environmental diversity of bacterial communities (Lozupone and Knight, 2008).

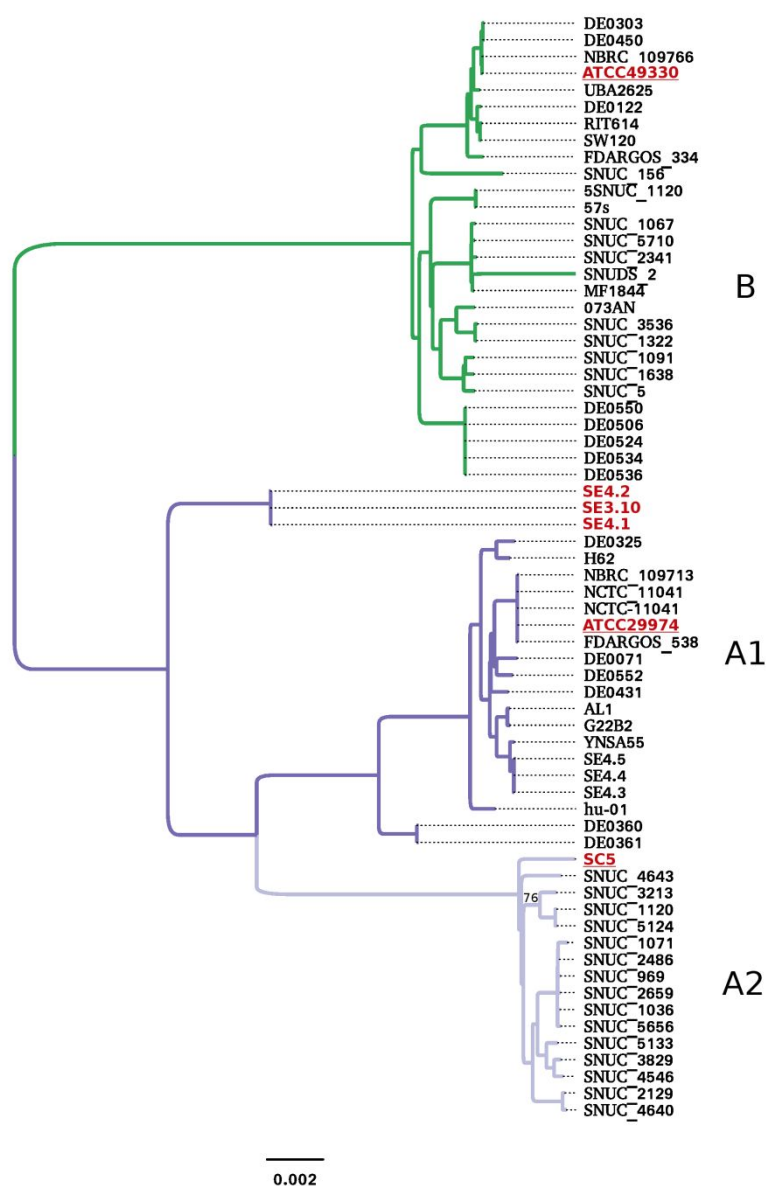


Fig. 4

Molecular discrimination of A1, A2 and B strains

Interestingly, analyses of core genome composition between the 3 SC groups proposed here, identify 10, 8, and 6 genes which are universally present in the genomes of A1, A2 and B strains respectively, and consistently absent from genomes of the other groups. These genes (table 2) are not adjacent in the genomes suggesting that they do not represent operons. PCR assays based on the presence/absence profiles of these genes might be used to discriminate between members of the A1, A2 and B groups.

Table 2.

A1, A2 and B strain-specific genes. The protein ID and the relevant annotation is shown for each gene.

Group	Accession	Annotation
B	WP_073342256.1	NAD(P)/FAD-dependent oxidoreductase
B	WP_073344697.1	DUF4352 domain-containing protein
B	WP_046206585.1	DUF4064 domain-containing protein
B	WP_073344420.1	ABC transporter substrate-binding protein
B	WP_073345446.1	NAD(P)H-dependent oxidoreductase
B	WP_103161674.1	Anion permease
B	WP_103161408.1	Glucose 1-dehydrogenase
B	WP_073343781.1	Energy-coupling factor transporter...
B	WP_073341550.1	Amino acid ABC transporter ATP-binding
B	WP_073342088.1	Thymidylate synthase
A1	WP_103211109.1	ABC-F family ATP-binding cassette domain-containing protein
A1	WP_019468720.1	tRNA (N6-isopentenyl adenosine(37)-C2)-methyltransferase MiaB
A1	WP_019468192.1	MFS transporter
A1	WP_040030451.1	Aldehyde dehydrogenase family protein
A1	WP_019468481.1	Trigger factor
A1	WP_103211478.1	M15 family metalloproteinase
A1	WP_040030229.1	Class 1b ribonucleoside-diphosphate reductase
A1	WP_019468907.1	Hypothetical protein
A2	WP_107523479.1	Orotidine-5'-phosphate decarboxylase
A2	WP_107505199.1	BtrH N-terminal domain-containing protein
A2	WP_019468295.1	Uracil phosphoribosyltransferase
A2	WP_181187692.1	Hypothetical protein
A2	WP_107384484.1	Arsenate reductase
A2	WP_107384163.1	LytTR family transcriptional regulator

Discussion

Difficulties in the taxonomic assignment of a novel *Staphylococcus* isolate based on well-established genome identity metrics (ANIb and DDH), prompted us to perform an extensive phylogenomic analysis of the *Staphylococcus cohnii* (SC) species complex, based on all (65) currently available draft genome assemblies. Analyses based on genomic identity levels, core and accessory genome size, gene content and phylogenetic approaches consistently and strongly suggest that the SC species complex, as currently defined is composed of at least 3 distinct groups, and advocate a revision of the current phylogenetic classification of SC.

Based on analyses presented in this study, and on the current guidelines and best practices for the classification of bacterial species (Varghese et al. 2015), we propose that the 2 SC subspecies as described by (Schleifer and Kloos, 1975), should be instead regarded as two distinct species: *Staphylococcus cohnii* (SC), corresponding to groups A1 and A2 in this work (fig. 2), and *Staphylococcus urealitycus* (SU), corresponding to group B. Additionally, the revised classification of SC should include 2 subspecies: *Staphylococcus cohnii* subsp. *cohnii* (SCC), corresponding to group A1, and *Staphylococcus cohnii* subsp. *barensis* (SCB), corresponding to group A2 and

including our novel isolate. Importantly, although the three different clades were not associated with any evident morphological or phenotypic difference, possible criteria for an effective discrimination of SU, SCC and SCB are proposed in the current study. These include the application of standard approaches based either on DDH and ANIb, two genome-wide similarity metrics, which are considered a reference standard for delineation of bacterial species. Moreover, according to our analyses a limited, but consistent number of lineage specific genes is observed in each of SU, SCC and SCB. In principle, simple tests based on targeted PCR resequencing of these genes could be used to develop a highly effective molecular assays for the discrimination of both species and subspecies proposed here.

While increased future environmental sampling may help overcome any ascertainment bias inherent in the available cohort, and to resolve the uncertainty of the phylogenetic placement of the three isolates from rice seeds, our study demonstrates a wider host range for SCC than previously hypothesized (Kloos & Wolfshohl 1983, 1991), and includes isolates from plants (rice seed) and vegetable liquid food (soy sauce) (table 3). Conversely isolates of SCB strains displayed a narrower hosts range, as 15 out of 16 strains were isolated from *Bos taurus* intramammary infections (bovine mastitis) and one (the novel strain described in this study) isolated from a disused biological cabinet. Intriguingly we observe that isolates of SCB also show a more compact genome and a significant reduction in gene content, compared to SCC a consideration that might at least in part explain its more reduced/specialized host range. The apparent host range of SU, based on the source of isolation as reported in the metadata associated with Genbank genome submissions, is consistent with previous reports with members of this candidate species isolated both from humans (e.g. skin, blood, catheter) and other animals, including ducks, dogs, cows and goats (table 3).

Table 3.

Staphylococcus cohnii genome metadata ^{a, b}

Strain	Isolation source	Assembly	Biosample	Bioproject	Geographic location	Current rank	Proposed rank
SC5	Biological Safety Cabinet	-	SAMN14142771	PRJNA607668	Bari, Italy	Sc	SCB
SNUC 2659	<i>Bos taurus</i> bovine mastitis	GCA_003035875.1	SAMN06172961	PRJNA342349	Quebec, Canada	Sc	SCB
SNUC 5656	<i>Bos taurus</i> bovine mastitis	GCA_003035785.1	SAMN06172970	PRJNA342349	Quebec, Canada	Sc	SCB
SNUC 5133	<i>Bos taurus</i> bovine mastitis	GCA_003577875.1	SAMN06172969	PRJNA342349	Quebec, Canada	Sc	SCB
SNUC 1036	<i>Bos taurus</i> bovine mastitis	GCA_003035975.1	SAMN06172951	PRJNA342349	Ontario, Canada	Sc	SCB
SNUC 1120	<i>Bos taurus</i> bovine mastitis	GCA_003035965.1	SAMN06172955	PRJNA342349	Quebec, Canada	Sc	SCB
SNUC 969	<i>Bos taurus</i> bovine mastitis	GCA_003039995.1	SAMN06172950	PRJNA342349	Quebec, Canada	Sc	SCB
SNUC 4643	<i>Bos taurus</i> bovine mastitis	GCA_003577905.1	SAMN06172967	PRJNA342349	Ontario, Canada	Sc	SCB
SNUC 1071	<i>Bos taurus</i> bovine mastitis	GCA_003578125.1	SAMN06172953	PRJNA342349	Ontario, Canada	Sc	SCB
SNUC 2129	<i>Bos taurus</i> bovine mastitis	GCA_003039915.1	SAMN06172958	PRJNA342349	Quebec, Canada	Sc	SCB
SNUC 5124	<i>Bos taurus</i> bovine mastitis	GCA_003035485.1	SAMN06172968	PRJNA342349	Quebec, Canada	Sc	SCB
SNUC 3213	<i>Bos taurus</i> bovine mastitis	GCA_003577955.1	SAMN06172962	PRJNA342349	Ontario, Canada	Sc	SCB
SNUC 2486	<i>Bos taurus</i> bovine mastitis	GCA_003035905.1	SAMN06172960	PRJNA342349	Atlantic, Canada	Sc	SCB
SNUC 3829	<i>Bos taurus</i> bovine mastitis	GCA_003035865.1	SAMN06172964	PRJNA342349	Ontario, Canada	Sc	SCB
SNUC 4546	<i>Bos taurus</i> bovine mastitis	GCA_003577915.1	SAMN06172965	PRJNA342349	Ontario, Canada	Sc	SCB
SNUC 4640	<i>Bos taurus</i> bovine mastitis	GCA_003035505.1	SAMN06172966	PRJNA342349	Ontario, Canada	Sc	SCB
FDARGOS_538	Human clinical isolate	GCA_003956025.1	SAMN10163250	PRJNA231221	Missing	Sc	SCC
SE4.4	Rice seed	GCA_001876785.1	SAMN03097241	PRJNA263233	India	Sc	SCC
SE4.3	Rice seed	GCA_001876755.1	SAMN03097240	PRJNA263233	India	Sc	SCC
ATCC29974	Human isolate	GCA_900240165.1	SAMEA104410613	PRJEB22856	Liverpool, UK	Sc	SCC
SE4.5	Rice seed	GCA_001876805.1	SAMN03097242	PRJNA263234	India	Sc	SCC
NCTC 11041	Human skin	GCA_002902365.1	SAMN06177162	PRJNA339206	London UK	Sc	SCC
NCTC 11041	Human skin	GCA_900458255.1	SAMEA3871778	PRJEB6403	USA	Sc	SCC
G22B2	Human gall bladder	GCA_000981215.1	SAMN03352186	PRJNA275680	Chandigarh, India	Sc	SCC
hu-01	Human skin swab sample	GCA_000513495.2	SAMN02388844	PRJNA225658	Hangzhou, China	Sc	SCC
AL1	Soy sauce (plant food)	GCA_000292305.1	SAMN02471867	PRJNA171726	Malaysia	Sc	SCC
DE0361	Environmental	GCA_007673385.1	SAMN11792521	PRJNA543692	Durham, North Carolina, USA	Sc	SCC
DE0431	Environmental	GCA_007668065.1	SAMN11792591	PRJNA543692	Durham, North Carolina, USA	Sc	SCC
DE0071	Environmental	GCA_007679825.1	SAMN11792231	PRJNA543692	Durham, North Carolina, USA	Sc	SCC
DE0552	Environmental	GCA_007666185.1	SAMN11792712	PRJNA543692	Durham, North Carolina, USA	Sc	SCC
DE0325	Environmental	GCA_008764065.1	SAMN11792485	PRJNA543692	Durham, North Carolina, USA	Sc	SCC
NBRC 109713	Human skin	GCA_007992675.1	SAMD00172682	PRJDB1638	Missing	Sc	SCC
DE0360	Environmental	GCA_007673395.1	SAMN11792520	PRJNA543692	Durham, North Carolina, USA	Sc	SCC
YNSA55	Human	GCA_005861955.1	SAMN11775280	PRJNA543691	Yunnan, China	Sc	SCC
H62	Environmental (air)	GCA_001650645.1	SAMN04591361	PRJNA316869	California, USA	Sc	SCC
SE4.1	Rice seed	GCA_001876705.1	SAMN03097238	PRJNA263229	India	Sc	SCC
SE4.2	Rice seed	GCA_001876735.1	SAMN03097239	PRJNA263230	India	Sc	SCC
SE3.10	Rice seed	GCA_001876725.1	SAMN03097237	PRJNA263228	India	Sc	SCC
FDARGOS_334	Human peripheral blood	GCA_002984565.1	SAMN06173347	PRJNA231221	Maryland USA	Sc	SU
532	Human catheter	GCA_000972575.1	SAMN03449104	PRJNA279286	Nuevo Leon, Monterrey, Mexico	Sc	SU
57	Human blood	GCA_000972565.1	SAMN03449103	PRJNA279286	Nuevo Leon, Monterrey, Mexico	Scu	SU
DSM6718-ATCC49330	Human skin	GCA_002902235.1	SAMN05977987	PRJNA339206	Braunschweig, Germany	Scu	SU
SW120	<i>Canis lupus</i> ear swab	GCA_001896245.1	SAMN06043532	PRJNA354224	Ballarat, Australia	Scu	SU
RIT614	Mobile phone (urban biome)	GCA_003725375.1	SAMN10392928	PRJNA504471	USA	Sc	SU
MF1844	Poultry processing equipment	GCA_001651275.1	SAMN04479463	PRJNA311173	Norway	Sc	SU
073AN	Goat perineal	GCA_900097955.1	SAMEA3109313	PRJEB2655	Tanzania, Africa	Scu/Sc	SU
SNUC 156	<i>Bos taurus</i> bovine mastitis	GCA_003035945.1	SAMN06172949	PRJNA342349	Alberta, Canada	Sc	SU
SNUC 1322	<i>Bos taurus</i> bovine mastitis	GCA_003039955.1	SAMN06172956	PRJNA342349	Ontario, Canada	Sc	SU
SNUC 5	<i>Bos taurus</i> bovine mastitis	GCA_003036005.1	SAMN06172948	PRJNA342349	Alberta, Canada	Sc	SU
SNUC 5710	<i>Bos taurus</i> bovine mastitis	GCA_003035835.1	SAMN06172971	PRJNA342349	Quebec, Canada	Sc	SU
SNUC 1091	<i>Bos taurus</i> bovine mastitis	GCA_003039975.1	SAMN06172954	PRJNA342349	Ontario, Canada	Sc	SU
SNUC 2341	<i>Bos taurus</i> bovine mastitis	GCA_003035925.1	SAMN06172959	PRJNA342349	Atlantic, Canada	Sc	SU
SNUC 1638	<i>Bos taurus</i> bovine mastitis	GCA_003039935.1	SAMN06172957	PRJNA342349	Ontario, Canada	Sc	SU
SNUC 3536	<i>Bos taurus</i> bovine mastitis	GCA_003577935.1	SAMN06172963	PRJNA342349	Ontario, Canada	Sc	SU
DE0122	Environmental	GCA_007679195.1	SAMN11792282	PRJNA543692	Durham, North Carolina, USA	Sc	SU
DE0450	Environmental	GCA_007667835.1	SAMN11792610	PRJNA543692	Durham, North Carolina, USA	Sc	SU
DE0303	Environmental	GCA_007674235.1	SAMN11792463	PRJNA543692	Durham, North Carolina, USA	Sc	SU
NBRC 109766	Human skin	GCA_007992755.1	SAMD00172686	PRJDB6345	Missing	Scu	SU
DE0536	Environmental	GCA_007666405.1	SAMN11792696	PRJNA543692	Durham, North Carolina, USA	Sc	SU
DE0534	Environmental	GCA_007666455.1	SAMN11792694	PRJNA543692	Durham, North Carolina, USA	Sc	SU
DE0524	Environmental	GCA_007666625.1	SAMN11792684	PRJNA543692	Durham, North Carolina, USA	Sc	SU
DE0550	Environmental	GCA_008763915.1	SAMN11792710	PRJNA543692	Durham, North Carolina, USA	Sc	SU
DE0506	Environmental	GCA_007666955.1	SAMN11792666	PRJNA543692	Durham, North Carolina, USA	Sc	SU
SNUDS 2	Duck brain	GCA_900240165.1	SAMN06286343	PRJNA369449	Seul, Sud Korea	Scu	SU
UBA2625	Environmental (metal)	GCA_002359925.1	SAMN06452948	PRJNA348753	New York City, USA	Sc	SU
SNUC 1067	<i>Bos taurus</i> bovine mastitis	GCA_003578165.1	SAMN06172952	PRJNA342349	Ontario, Canada	Sc	SU

^a Metadata as reported in Genbank. ^b Colors were assigned according to the 3 cluster memberships as in Figure 1, A1 = dark purple; A2 = light purple; B = green. Current ranks: Sc, *Staphylococcus cohnii*; Sc, *Staphylococcus cohnii* subsp. *cohnii*; Scu, *Staphylococcus cohnii* subsp. *urealyticus*. Proposed ranks: SCB, *Staphylococcus cohnii* subsp. *Barensis*; SCC, *Staphylococcus cohnii* subsp. *cohnii*; SU, *Staphylococcus urealyticus*. In bold, SC5 and type strains (ATCC 29974 and ATCC49330).

Consistent with the problems encountered in the taxonomic classification of our novel SCB isolate, we observed that taxonomic classifications of several SC isolates deposited in Genbank are incomplete or discordant with our analyses. For example, strain 532 (Mendoza-Olazarán et al. 2017) is assigned to SCC in the NCBI biosample, but is likely a member of SCU. Furthermore, strain 073AN (NCBI draft genome A.N.: FMPF00000000.1), isolated in

Tanzania (Africa) from a goat, is labelled as SCC in the associated Biosample and Bioproject submissions (SAMEA3109313 and ERS576551) (INSDC, last update 2016-09-28, <https://www.ncbi.nlm.nih.gov/biosample/SAMEA3109313/>) although both the original publication (Seni et al. 2019) and our own analyses indicate that it is SCU. These misclassifications are likely consequences of problems in the initial classification of the isolates, arising from ambiguities in the delineation of the SC species complex. Consistent with this hypothesis, we observe that in many cases the classification of SC strains is limited to species level.

Misclassified sequences, errors in initial taxonomic classification and annotation can occur and be propagated - resulting in the unintentional misclassification of bacterial strains (Federhen 2015). Accordingly, we believe that the approach and criteria presented here, may be of general interest and could be applied on a larger scale for the resolution of complex/conflicting taxonomic assignments.

Data availability

The SC5 draft genome is available in NCBI under the accession number JAALCY000000000, BioSample accession number SAMN14142771 and BioProject number ID PRJNA607668.

Data are available for peer review to <https://drive.google.com/drive/folders/1DFraH-P1c6bqTHVxBnuR4ozOEFskwRdY?usp=sharing>.

Supplementary Material

Supplementary data are available at Genome Biology and Evolution online.

Acknowledgments

This work was supported by INMARE (H2020-BG-2014-2, GA 634486), EMBRIC (H2020-INFRADEV-1-2014-1, GA 654008), EXCELERATE (H2020-INFRADEV-1-2015-1, GA676559), the Molecular Biodiversity Laboratory (Life Watch, Italy), and the Italian Node of ELIXIR. The funder had no role in the study design, data collection and interpretation, or the decision to submit the work for publication.

Literature Cited

Adeyemi AI, Sulaiman AA, Solomon BB, Chinedu OA, Victor IA. 2010. Bacterial Bloodstream Infections in HIV-infected Adults Attending a Lagos Teaching Hospital. *J Health Popul Nutr.* 28:318–326.

- Afgan E et al. 2018. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.* 46: W537–W544. doi: 10.1093/nar/gky379.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol.* 215:403–410. doi: 10.1016/S0022-2836(05)80360-2.
- Argemi X, Hansmann Y, Prola K, Prévost G. 2019. Coagulase-Negative Staphylococci Pathogenomics. *Int J Mol Sci.* 20:1215. doi: 10.3390/ijms20051215.
- Auch AF, Klenk HP, Göker M. 2010. Standard operating procedure for calculating genome-to-genome distances based on high-scoring segment pairs. *Stand Genomic Sci.* 2(1):142-148. doi:10.4056/sigs.541628
- Bankevich A et al. 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J Comput Biol.* 19:455–477. doi: 10.1089/cmb.2012.0021.
- Basaglia G, Moras L, Bearz A, Scalone S, Paoli PD. 2003. *Staphylococcus cohnii* septicaemia in a patient with colon cancer. *J Med Microbiol.* 52:101–102. doi: 10.1099/jmm.0.05002-0.
- Bean DC, Wigmore SM, Wareham DW. 2017. Draft Genome Sequence of *Staphylococcus cohnii* subsp. *urealyticus* Isolated from a Healthy Dog. *Genome Announc.* 5. doi: 10.1128/genomeA.01628-16.
- Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics.* 30:2114–2120. doi: 10.1093/bioinformatics/btu170.
- Camacho C., Coulouris G., Avagyan V., Ma N., Papadopoulos J., Bealer K., & Madden T.L. 2008 "BLAST+: architecture and applications." *BMC Bioinformatics* 10:421
- Castresana J. 2000. Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. *Mol Biol Evol.* 17:540–552. doi: 10.1093/oxfordjournals.molbev.a026334.
- Chiara M et al. 2018. A-GAME: improving the assembly of pooled functional metagenomics sequence data. *BMC Genomics.* 19:44. doi: 10.1186/s12864-017-4369-z.
- Colston SM et al. 2014. Bioinformatic genome comparisons for taxonomic and phylogenetic assignments using *Aeromonas* as a test case. *mBio.* 5: e02136. doi: 10.1128/mBio.02136-14.
- Crossley KB, Jefferson KK, Archer GL, Jr VGF, eds. 2009. *Staphylococci in Human Disease*. 2 edition. Wiley-Blackwell: Chichester, West Sussex; Hoboken, NJ.
- Darriba D, Taboada GL, Doallo R, Posada D. 2011. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics.* 27:1164–1165. doi: 10.1093/bioinformatics/btr088.
- David MD, Elliott T. 2015. Coagulase-negative staphylococci. *Br J Hosp Med (Lond).* 76:C126-128. doi: 10.12968/hmed.2015.76.8.C126.
- Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32:1792–1797. doi: 10.1093/nar/gkh340.
- Federhen S. 2015. Type material in the NCBI Taxonomy Database. *Nucleic Acids Res.* 43: D1086-1098. doi: 10.1093/nar/gku1127.
- Garrido-Sanz D et al. 2016. Correction: Genomic and Genetic Diversity within the *Pseudomonas fluorescens* Complex. *PLoS ONE.* 11: e0153733. doi: 10.1371/journal.pone.0153733.
- Guindon S, Delsuc F, Dufayard J-F, Gascuel O. 2009. Estimating Maximum Likelihood Phylogenies with PhyML. In: *Bioinformatics for DNA Sequence Analysis*. Posada, D, editor. *Methods Molecular Biology Humana Press*: Totowa, NJ pp. 113–137. doi: 10.1007/978-1-59745-251-9_6.
- Kloos WE, Wolfshohl JF. 1983. Deoxyribonucleotide sequence divergence between *Staphylococcus cohnii* subspecies populations living on primate skin. *Curr Microbi.* 8:115–121. doi: 10.1007/BF01566968.

- Kloos WE, Wolfshohl JF. 1991. *Staphylococcus cohnii* subspecies: *Staphylococcus cohnii* subsp. *cohnii* subsp. *nov.* and *Staphylococcus cohnii* subsp. *urealyticum* subsp. *nov.* Int J Syst Bacteriol. 41:284–289. doi: 10.1099/00207713-41-2-284.
- Larsen J, Andersen PS, Winstel V, Peschel A. 2017. *Staphylococcus aureus* CC395 harbours a novel composite staphylococcal cassette chromosome mec element. J Antimicrob Chemother. 72:1002–1005. doi: 10.1093/jac/dkw544.
- Lavecchia A et al. 2018. Draft Genome Sequences of Three Novel *Staphylococcus arlettae* Strains Isolated from a Disused Biological Safety Cabinet. Microbiol Resour Announc. 7. doi: 10.1128/MRA.01012-18.
- Lozupone CA, Knight R. Species divergence and the measurement of microbial diversity. 2008. FEMS Microbiol Rev. 32(4):557–578. doi:10.1111/j.1574-6976.2008.00111.x)
- Mastroianni A, Coronado O, Nanetti A, Manfredi R, Chiodo F. 1996. *Staphylococcus cohnii*: an unusual cause of primary septic arthritis in a patient with AIDS. Clin Infect Dis. 23:1312–1313. doi: 10.1093/clinids/23.6.1312.
- Meier-Kolthoff JP, Auch AF, Klenk H-P, Göker M. 2013. Genome sequence-based species delimitation with confidence intervals and improved distance functions. BMC Bioinformatics. 14:60. doi: 10.1186/1471-2105-14-60.
- Mendoza-Olazarán S et al. 2017. Draft genome sequences of two opportunistic pathogenic strains of *Staphylococcus cohnii* isolated from human patients. Stand Genomic Sci. 12. doi: 10.1186/s40793-017-0263-1.
- Moawad AA et al. 2019. Evolution of Antibiotic Resistance of Coagulase-Negative Staphylococci Isolated from Healthy Turkeys in Egypt: First Report of Linezolid Resistance. Microorganisms. 7. doi: 10.3390/microorganisms7100476.
- Okudera H, Kobayashi S, Hongo K, Mizuno M. 1991. Fatal meningitis due to *Staphylococcus cohnii*. Case report. Neurosurg Rev. 14:235–236. doi: 10.1007/BF00310665.
- Otto M. 2013. Coagulase-negative staphylococci as reservoirs of genes facilitating MRSA infection. Bioessays. 35:4–11. doi: 10.1002/bies.201200112.
- Otto M. 2008. Staphylococcal Biofilms. Curr Top Microbiol Immunol. 322:207–228.
- Richter M, Rosselló-Móra R. 2009. Shifting the genomic gold standard for the prokaryotic species definition. Proc. Natl. Acad. Sci. U.S.A. 106:19126–19131. doi: 10.1073/pnas.0906412106.
- Schleifer KH, Kloos WE. 1975. Isolation and Characterization of Staphylococci from Human Skin I. Amended Descriptions of *Staphylococcus epidermidis* and *Staphylococcus saprophyticus* and Descriptions of Three New Species: *Staphylococcus cohnii*, *Staphylococcus haemolyticus*, and *Staphylococcus xylosus*. Int Syst Evol Micr. 25:50–61. doi: 10.1099/00207713-25-1-50.
- Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 30:2068–2069. doi: 10.1093/bioinformatics/btu153.
- Seni J et al. 2019. Draft genome sequence of a multidrug-resistant caprine isolate of *Staphylococcus cohnii* subsp. *urealyticus* from Tanzania encoding ermB, tet(K), dfrG, fusF and fosD. J Glob Antimicrob Resist. 18:163–165. doi: 10.1016/j.jgar.2019.07.006.
- Sneath PH. 1992. Correction of orthography of epithets in Pasteurella and some problems with recommendations on latinization. Int J Syst Bacteriol. 42:658–659. doi: 10.1099/00207713-42-4-658.
- Song Y et al. 2017. cfr-mediated linezolid-resistant clinical isolates of methicillin-resistant coagulase-negative staphylococci from China. J Glob Antimicrob Resist. 8:1–5. doi: 10.1016/j.jgar.2016.09.008.
- Szczuka E, Jabłońska L, Kaznowski A. 2016. Coagulase-negative staphylococci: pathogenesis, occurrence of antibiotic resistance genes and in vitro effects of antimicrobial agents on biofilm-growing bacteria. J Med Microbiol. 65:1405–1413. doi: 10.1099/jmm.0.000372.

- Szewczyk EM, Piotrowski A, Różalska M. 2000. Predominant staphylococci in the intensive care unit of a paediatric hospital. *J Hosp Infect.* 45:145–154. doi: 10.1053/jhin.1999.0754.
- Szewczyk EM, Różalska M, Cieślowski T, Nowak T. 2004. Plasmids of *Staphylococcus cohnii* isolated from the intensive-care unit. *Folia Microbiol.* 49:123–131. doi: 10.1007/BF02931385.
- Varghese NJ et al. 2015. Microbial species delineation using whole genome sequences. *Nucleic Acids Res.* 43:6761–6771. doi: 10.1093/nar/gkv657.
- Waldon E, Sobiś-Glinkowska M, Szewczyk EM. 2002. Evaluation of selected features of *Staphylococcus cohnii* enabling colonization of humans. *Folia Microbiol. (Praha).* 47:565–571. doi: 10.1007/BF02818799.
- Whelan S, Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.* 18:691–699. doi: 10.1093/oxfordjournals.molbev.a003851.
- Winstel V et al. 2013. Wall teichoic acid structure governs horizontal gene transfer between major bacterial pathogens. *Nat Commun.* 4:2345. doi: 10.1038/ncomms3345.
- Yamashita S, Yonemura K, Sugimoto R, Tokunaga M, Uchino M. 2005. *Staphylococcus cohnii* as a cause of multiple brain abscesses in Weber–Christian disease. *J Neurol Sci.* 238:97–100. doi: 10.1016/j.jns.2005.06.009.
- Yong YY, Dykes GA, Choo WS. 2019. Biofilm formation by staphylococci in health-related environments and recent reports on their control using natural compounds. *Crit Rev Microbiol.* 45:201–222. doi: 10.1080/1040841X.2019.1573802.
- Zhang H, Ma Y, Liu P, Li X. 2016. Multidrug resistance operon *emrAB* contributes for chromate and ampicillin co-resistance in a *Staphylococcus* strain isolated from refinery polluted river bank. *Springerplus.* 5:1648. doi: 10.1186/s40064-016-3253-7.
- Zhang J, Kobert K, Flouri T, Stamatakis A. 2014. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics.* 30:614–620. doi: 10.1093/bioinformatics/btt593.

Figure Legends:

Fig. 1: Heatmap of ANIb (Average nucleotide identity based on BLAST) between genomes of *Staphylococcus cohnii* isolates. ANIb values are represented using a gray scale color map, with darker colors indicating higher levels of identity, according to the scale represented on the top. Strain identifiers are indicated on the rows. The panel on the left indicates cluster memberships, according to the following color codes: green = B, dark purple = A1, and light purple = A2. Columns and row dendrograms are used to group SC strains based on patterns of genome identity profiles. The novel SC5 isolate and the two type strains SCC ATCC29974 and SCU ATCC49330 are highlighted in red and underlined. The SE4.1, SE 4.2 and SE3.10 strains, which are also discussed in the text are highlighted in red.

Fig. 2: Heatmap of gene absence/presence profiles of *Staphylococcus cohnii* isolates. Similarity of gene absence/presence profiles were estimated by computing pairwise Pearson correlation values between all the 66

genomes considered in the study. Pearson correlation coefficients are represented using a gray scale color map. Darker colors indicated higher correlation (similarity) of gene absence/presence profiles. Strain identifiers are indicated on the rows. The panel on the left is used to indicate cluster memberships, with the color codes defined in Figure 1. Similar to Figure 1, dendrograms are applied to the columns and rows to delineate groups of isolates with similar gene absence prevalence profiles. SC5 and the two type strains SCC ATCC29974 and SCU ATCC49330 are highlighted in red and underlined. The SE4.1, SE 4.2 and SE3.10 strains are highlighted in red.

Fig. 3: Core and accessory genome. A) Plot of core genome size in the *Staphylococcus cohnii* species complex and in the 3 distinct groups of SC (A1, A2, and B) delineated in this study. X axis = number of genomes. Y axis = number of genes. B) Plot of accessory genome size in *S. cohnii* and in the 3 distinct groups of SC identified by this study. X axis = number of genomes. Y axis = number of genes.

Fig. 4: Phylogenetic tree of *Staphylococcus cohnii* isolates based on concatenated alignment of 1468 core genes. Branch colors indicate the different groups identified in this study, according to the color code defined in Figure 1. Bootstrap values below 95 are reported on the corresponding branches. The SE4.1, SE 4.2 and SE3.10 strains are marked in red. SC5 isolate and the two type strains SCC ATCC29974 and SCU ATCC49330 are highlighted in red and underlined.

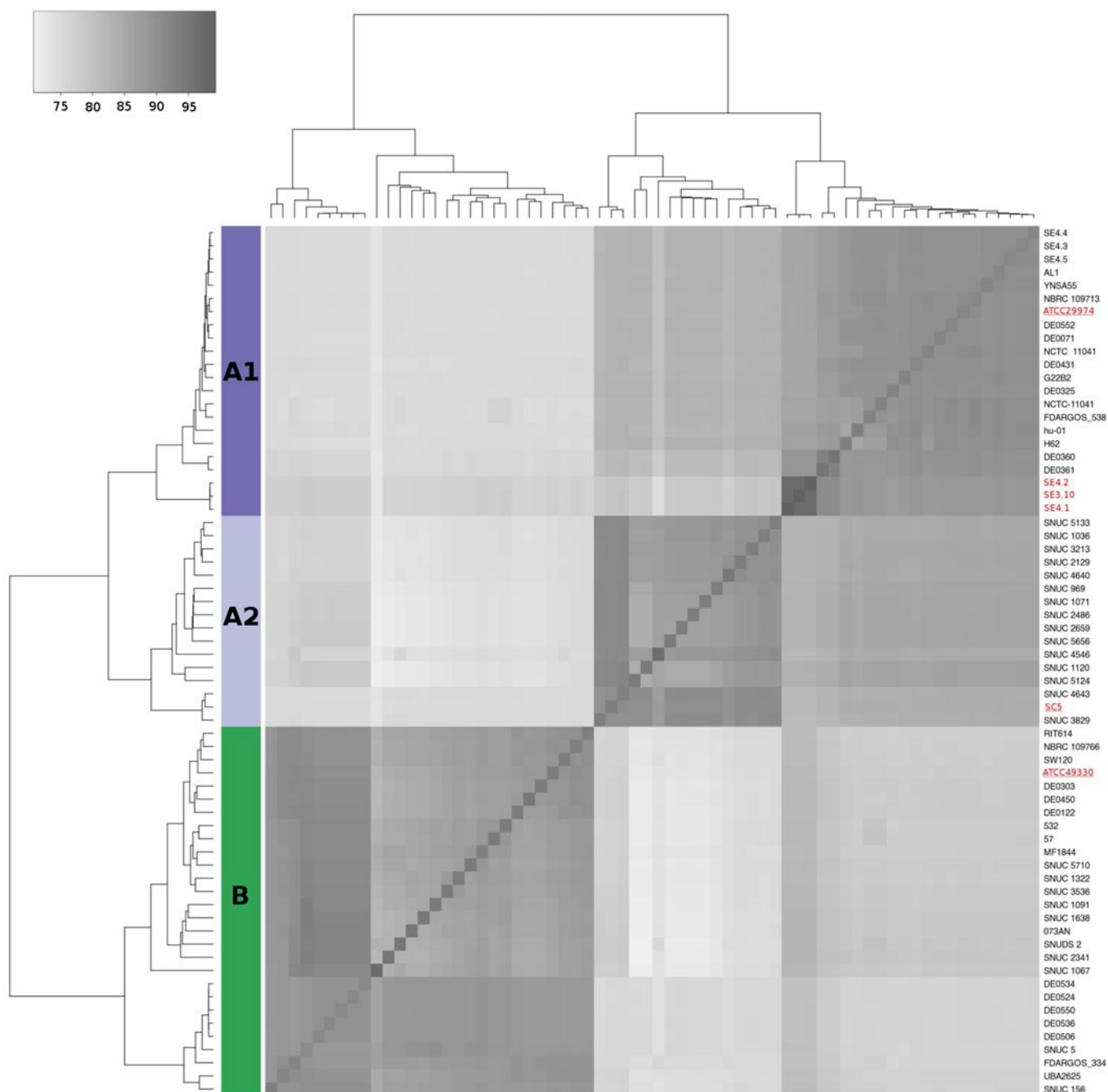


Figure 1: Heatmap of ANIb (Average Nucleotide Identity based on BLAST) between genomes of *Staphylococcus cohnii* isolates. ANIb values are represented using a gray scale color map, with darker colors indicating higher levels of identity, according to the scale represented on the top. Strain identifiers are indicated on the rows. The panel on the left indicates cluster memberships, according to the following color codes: green = B, dark purple = A1, and light purple = A2. Columns and row dendrograms are used to group SC strains based on patterns of genome identity profiles. The novel SC5 isolate and the two type strains SCC ATCC29974 and SCU ATCC49330 are highlighted in red and underlined. The SE4.1, SE4.2 and SE3.10 strains, which are also discussed in the text are highlighted in red.

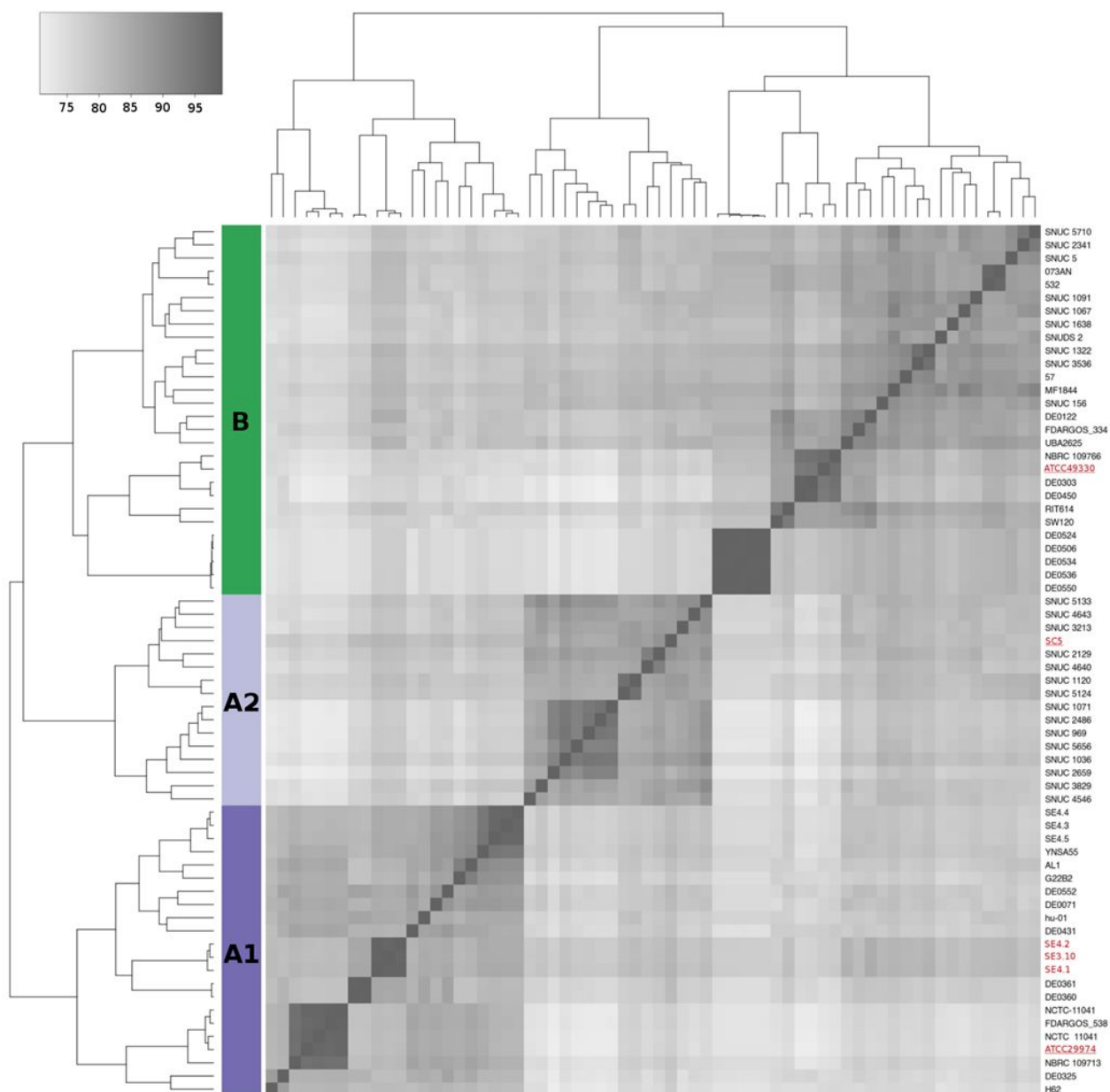


Figure 2: Heatmap of gene absence/presence profiles between *Staphylococcus cohnii* isolates. Similarity of gene absence/presence profiles were estimated by computing pairwise Pearson correlation values between all the 66 genomes considered in the study. Pearson correlation coefficients are represented using a gray scale color map. Darker colors indicated higher correlation (similarity) of gene absence/presence profiles. Strain identifiers are indicated on the rows. The panel on the left is used to indicate cluster memberships, with the color code defined in Figure 1. Similar to Figure 1, dendrograms are applied to the columns and rows to delineate groups of isolates with similar gene absence profiles. SC5 and the two type strains SCC ATCC29974 and SCU ATCC49330 are highlighted in red and underlined. The SE4.1, SE 4.2 and SE3.10 strains are highlighted in red.

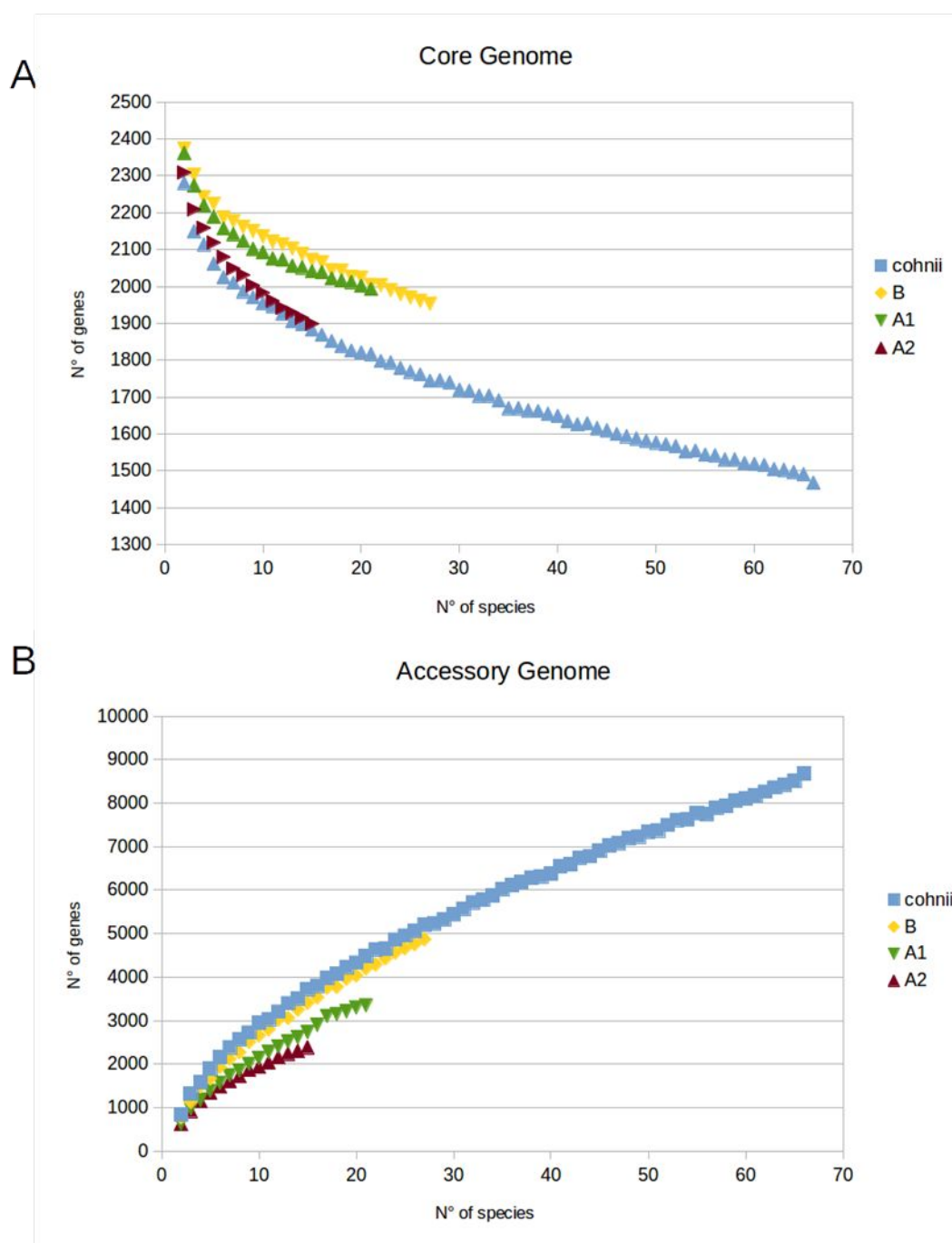


Figure 3: Core and accessory genome. A) Plot of core genome size in *Staphylococcus cohnii* species complex and in the 3 distinct groups of SC (A1, A2, and B) delineated in this study. X axis = number of genomes. Y axis = number of genes. B) Plot of accessory genome size in *Staphylococcus cohnii* species and in the 3 distinct groups of SC identified by this study. X axis = number of genomes. Y axis = number of genes.

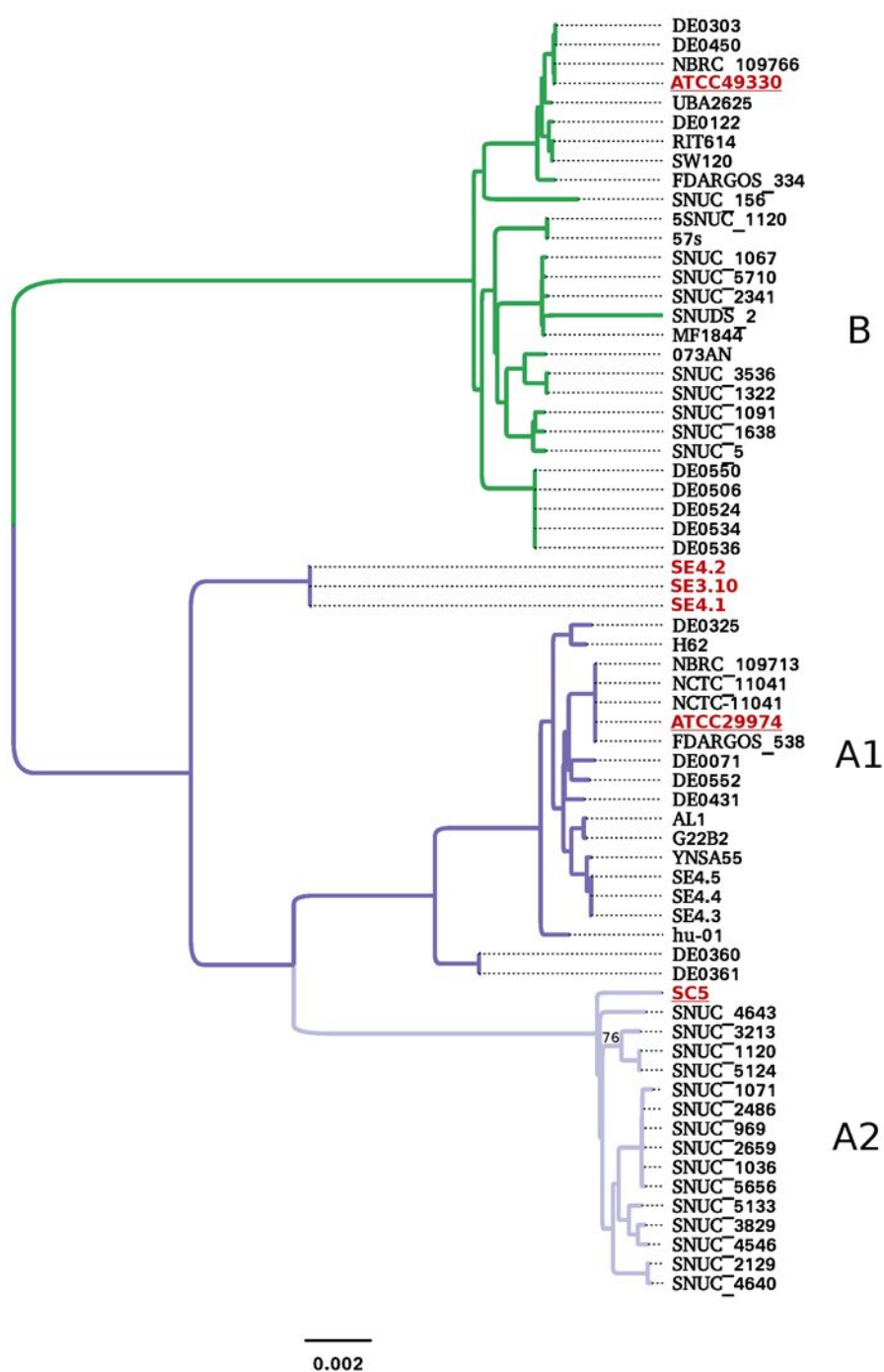


Figure 4: Phylogenetic tree of *Staphylococcus cohnii* isolates based on concatenated alignment of 1468 core genes. Branch colors indicate the different groups identified in this study, according to the color code defined in Figure 1. Bootstrap values below 95 are reported on the corresponding branches. The SE4.1, SE4.2 and SE3.10 strains are marked in red. SC5 isolate and the two type strains SCC ATCC29974 and SCU ATCC49330 are highlighted in red and underlined.

Table 1.

Main genome assembly features of SC5, SCC ATCC 29974 and SCU ATCC 49330 strains. DDH and ANIb values between strains.

Strain	Size (Mb)	GC (%)	Contigs	N50 (Kb)	Proteins	rRNAs	tRNAs	DDH (%) ^a			ANIb (%) ^b		
								SC5	SCC	SCU	SC5	SCC	SCU
SC5	2.62	32.2	34	806	2510	10	61	-	67.5	41.9	-	95.4	91.0
SCC ATCC 29974	2.71	32.6	83	114	2422	9	58	67.5	-	-	95.4	-	91.5
SCU ATCC 49330	2.67	32.5	223	26	2457	13	61	41.9	-	-	91.0	91.5	-

^a DDH, DNA-DNA hybridization (cut-off for species affiliation > 70%); ^b ANIb, Average Nucleotide Identity (cut-off for species affiliation > 96%). SC5, our isolate; SCC ATCC29974, *Staphylococcus cohnii* subsp. *cohnii* (type strain); SCU ATCC49330, *Staphylococcus cohnii* subsp. *urealyticus* (type strain).

Table 2.

A1, A2 and B strain-specific genes. The protein ID and the relevant annotation is shown for each gene.

Group	Accession	Annotation
B	WP_073342256.1	NAD(P)/FAD-dependent oxidoreductase
B	WP_073344697.1	DUF4352 domain-containing protein
B	WP_046206585.1	DUF4064 domain-containing protein
B	WP_073344420.1	ABC transporter substrate-binding protein
B	WP_073345446.1	NAD(P)H-dependent oxidoreductase
B	WP_103161674.1	Anion permease
B	WP_103161408.1	Glucose 1-dehydrogenase
B	WP_073343781.1	Energy-coupling factor transporter...
B	WP_073341550.1	Amino acid ABC transporter ATP-binding
B	WP_073342088.1	Thymidylate synthase
A1	WP_103211109.1	ABC-F family ATP-binding cassette domain-containing protein
A1	WP_019468720.1	tRNA (N6-isopentenyl adenosine(37)-C2)-methyltransferase MiaB
A1	WP_019468192.1	MFS transporter
A1	WP_040030451.1	Aldehyde dehydrogenase family protein
A1	WP_019468481.1	Trigger factor
A1	WP_103211478.1	M15 family metalloproteinase
A1	WP_040030229.1	Class 1b ribonucleoside-diphosphate reductase
A1	WP_019468907.1	Hypothetical protein
A2	WP_107523479.1	Orotidine-5'-phosphate decarboxylase
A2	WP_107505199.1	BtrH N-terminal domain-containing protein
A2	WP_019468295.1	Uracil phosphoribosyltransferase
A2	WP_181187692.1	Hypothetical protein
A2	WP_107384484.1	Arsenate reductase
A2	WP_107384163.1	LytTR family transcriptional regulator

Table 3. *Staphylococcus cohnii* genome metadata ^{a, b}

Strain	Isolation source	Assembly	Biosample	Bioproject	Geographic location	Current rank	Proposed rank
SC5	Biological Safety Cabinet	-	SAMN14142771	PRJNA607668	Bari, Italy	Sc	SCB
SNUC 2659	<i>Bos taurus</i> bovine mastitis	GCA_003035875.1	SAMN06172961	PRJNA342349	Quebec, Canada	Sc	SCB
SNUC 5656	<i>Bos taurus</i> bovine mastitis	GCA_003035785.1	SAMN06172970	PRJNA342349	Quebec, Canada	Sc	SCB
SNUC 5133	<i>Bos taurus</i> bovine mastitis	GCA_003577875.1	SAMN06172969	PRJNA342349	Quebec, Canada	Sc	SCB
SNUC 1036	<i>Bos taurus</i> bovine mastitis	GCA_003035975.1	SAMN06172951	PRJNA342349	Ontario, Canada	Sc	SCB
SNUC 1120	<i>Bos taurus</i> bovine mastitis	GCA_003035965.1	SAMN06172955	PRJNA342349	Quebec, Canada	Sc	SCB
SNUC 969	<i>Bos taurus</i> bovine mastitis	GCA_003039995.1	SAMN06172950	PRJNA342349	Quebec, Canada	Sc	SCB
SNUC 4643	<i>Bos taurus</i> bovine mastitis	GCA_003035905.1	SAMN06172967	PRJNA342349	Ontario, Canada	Sc	SCB
SNUC 1071	<i>Bos taurus</i> bovine mastitis	GCA_003578125.1	SAMN06172953	PRJNA342349	Ontario, Canada	Sc	SCB
SNUC 2129	<i>Bos taurus</i> bovine mastitis	GCA_003039915.1	SAMN06172958	PRJNA342349	Quebec, Canada	Sc	SCB
SNUC 5124	<i>Bos taurus</i> bovine mastitis	GCA_003035485.1	SAMN06172968	PRJNA342349	Quebec, Canada	Sc	SCB
SNUC 3213	<i>Bos taurus</i> bovine mastitis	GCA_003577955.1	SAMN06172962	PRJNA342349	Ontario, Canada	Sc	SCB
SNUC 2486	<i>Bos taurus</i> bovine mastitis	GCA_003035905.1	SAMN06172960	PRJNA342349	Atlantic, Canada	Sc	SCB
SNUC 3829	<i>Bos taurus</i> bovine mastitis	GCA_003035865.1	SAMN06172964	PRJNA342349	Ontario, Canada	Sc	SCB
SNUC 4546	<i>Bos taurus</i> bovine mastitis	GCA_003577915.1	SAMN06172965	PRJNA342349	Ontario, Canada	Sc	SCB
SNUC 4640	<i>Bos taurus</i> bovine mastitis	GCA_003035805.1	SAMN06172966	PRJNA342349	Ontario, Canada	Sc	SCB
FDARGOS_538	Human clinical isolate	GCA_003956025.1	SAMN10163250	PRJNA231221	Missing	Sc	SCB
SE4.4	Rice seed	GCA_001876785.1	SAMN03097241	PRJNA263233	India	Sc	SCB
SE4.3	Rice seed	GCA_001876755.1	SAMN03097240	PRJNA263231	India	Sc	SCB
ATCC29974	Human isolate	GCA_900240165.1	SAMEA104410613	PRJEB22856	Liverpool, UK	Sc	SCB
SE4.5	Rice seed	GCA_001876805.1	SAMN03097242	PRJNA263234	India	Sc	SCB
NCTC_11041	Human skin	GCA_002902365.1	SAMN06177162	PRJNA339206	London UK	Sc	SCB
NCTC_11041	Human skin	GCA_900458255.1	SAMEA3871778	PRJEB6403	USA	Sc	SCB
G22B2	Human gall bladder	GCA_000981215.1	SAMN03352186	PRJNA275680	Chandigarh, India	Sc	SCB
hu-01	Human skin swab sample	GCA_000513495.2	SAMN02388844	PRJNA225658	Hangzhou, China	Sc	SCB
AL1	Soy sauce (plant food)	GCA_000292305.1	SAMN02471867	PRJNA171726	Malaysia	Sc	SCB
DE0361	Environmental	GCA_007673385.1	SAMN11792521	PRJNA543692	Durham, North Carolina, USA	Sc	SCB
DE0431	Environmental	GCA_007668065.1	SAMN11792591	PRJNA543692	Durham, North Carolina, USA	Sc	SCB
DE0071	Environmental	GCA_007679825.1	SAMN11792231	PRJNA543692	Durham, North Carolina, USA	Sc	SCB
DE0552	Environmental	GCA_007666185.1	SAMN11792712	PRJNA543692	Durham, North Carolina, USA	Sc	SCB
DE0325	Environmental	GCA_008764065.1	SAMN11792485	PRJNA543692	Durham, North Carolina, USA	Sc	SCB
NBRC_109713	Human skin	GCA_007992675.1	SAMD00172682	PRJDB1638	Missing	Sc	SCB
DE0360	Environmental	GCA_007673395.1	SAMN11792520	PRJNA543692	Durham, North Carolina, USA	Sc	SCB
YNSA55	Human	GCA_005861955.1	SAMN11775280	PRJNA543691	Yunnan, China	Sc	SCB
H62	Environmental (air)	GCA_001650645.1	SAMN04591361	PRJNA316869	California, USA	Sc	SCB
SE4.1	Rice seed	GCA_001876705.1	SAMN03097238	PRJNA263229	India	Sc	SCB
SE4.2	Rice seed	GCA_001876735.1	SAMN03097239	PRJNA263230	India	Sc	SCB
SE3.10	Rice seed	GCA_001876725.1	SAMN03097237	PRJNA263228	India	Sc	SCB
FDARGOS_334	Human peripheral blood	GCA_002984565.1	SAMN06173347	PRJNA231221	Maryland USA	Sc	SCB
532	Human catheter	GCA_000972575.1	SAMN03449104	PRJNA279286	Nuevo Leon, Monterrey, Mexico	Sc	SCB
57	Human blood	GCA_000972565.1	SAMN03449103	PRJNA279286	Nuevo Leon, Monterrey, Mexico	Sc	SCB
DSM6718=ATCC49330	Human skin	GCA_002902235.1	SAMN05977987	PRJNA339206	Braunschweig, Germany	Sc	SCB
SW120	<i>Canis lupus</i> ear swab	GCA_001896245.1	SAMN06043532	PRJNA354224	Ballarat, Australia	Sc	SCB
RIT614	Mobile phone (urban biome)	GCA_003725375.1	SAMN10392928	PRJNA504471	USA	Sc	SCB
MF1844	Poultry processing equipment	GCA_001651275.1	SAMN04479463	PRJNA311173	Norway	Sc	SCB
073AN	Goat perineal	GCA_900097955.1	SAMEA3109313	PRJEB2655	Tanzania, Africa	Scu/Sc	SCB
SNUC 156	<i>Bos taurus</i> bovine mastitis	GCA_003035945.1	SAMN06172949	PRJNA342349	Alberta, Canada	Sc	SCB
SNUC 1322	<i>Bos taurus</i> bovine mastitis	GCA_003039955.1	SAMN06172956	PRJNA342349	Ontario, Canada	Sc	SCB
SNUC 5	<i>Bos taurus</i> bovine mastitis	GCA_003036005.1	SAMN06172948	PRJNA342349	Alberta, Canada	Sc	SCB
SNUC 5710	<i>Bos taurus</i> bovine mastitis	GCA_003035835.1	SAMN06172971	PRJNA342349	Quebec, Canada	Sc	SCB
SNUC 1091	<i>Bos taurus</i> bovine mastitis	GCA_003039975.1	SAMN06172954	PRJNA342349	Ontario, Canada	Sc	SCB
SNUC 2341	<i>Bos taurus</i> bovine mastitis	GCA_003035925.1	SAMN06172959	PRJNA342349	Atlantic, Canada	Sc	SCB
SNUC 1638	<i>Bos taurus</i> bovine mastitis	GCA_003039935.1	SAMN06172957	PRJNA342349	Ontario, Canada	Sc	SCB
SNUC 3536	<i>Bos taurus</i> bovine mastitis	GCA_003577935.1	SAMN06172963	PRJNA342349	Ontario, Canada	Sc	SCB
DE0122	Environmental	GCA_007679195.1	SAMN11792282	PRJNA543692	Durham, North Carolina, USA	Sc	SCB
DE0450	Environmental	GCA_007667835.1	SAMN11792610	PRJNA543692	Durham, North Carolina, USA	Sc	SCB
DE0303	Environmental	GCA_007674235.1	SAMN11792463	PRJNA543692	Durham, North Carolina, USA	Sc	SCB
NBRC_109766	Human skin	GCA_007992755.1	SAMD00172686	PRJDB6345	Missing	Scu	SCB
DE0536	Environmental	GCA_007666405.1	SAMN11792696	PRJNA543692	Durham, North Carolina, USA	Sc	SCB
DE0534	Environmental	GCA_007666455.1	SAMN11792694	PRJNA543692	Durham, North Carolina, USA	Sc	SCB
DE0524	Environmental	GCA_007666625.1	SAMN11792684	PRJNA543692	Durham, North Carolina, USA	Sc	SCB
DE0550	Environmental	GCA_008763915.1	SAMN11792710	PRJNA543692	Durham, North Carolina, USA	Sc	SCB
DE0506	Environmental	GCA_007666955.1	SAMN11792666	PRJNA543692	Durham, North Carolina, USA	Sc	SCB
SNUDS 2	Duck brain	GCA_900240165.1	SAMN06286343	PRJNA369449	Seul, Sud Corea	Scu	SCB
UBA2625	Environmental (metal)	GCA_002359925.1	SAMN06452948	PRJNA348753	New York City, USA	Sc	SCB
SNUC 1067	<i>Bos taurus</i> bovine mastitis	GCA_003578165.1	SAMN06172952	PRJNA342349	Ontario, Canada	Sc	SCB

^a Metadata as reported in Genbank. ^b Colors were assigned according to the 3 cluster memberships as in Figure 1, A1 = dark purple; A2 = light purple; B = green. Sc, *Staphylococcus cohnii*; Scs, *Staphylococcus cohnii* subsp. *cohnii*; Scu, *Staphylococcus cohnii* subsp. *urealyticus*. SCB, *Staphylococcus cohnii* subsp. *barensis*; SCC, *Staphylococcus cohnii* subsp. *cohnii*; SU, *Staphylococcus urealyticus*. In bold, SC5 and type strains (ATCC29274 and ATCC49330).