Supporting Information for

Integrated genomic and functional analyses of human skin-associated *Staphylococcus* reveal extensive inter- and intra-species diversity

Payal Joglekar [a], Sean Conlan [a], Shih-Queen Lee-Lin [a], Clay Deming [a], Sara Saheb Kashaf [a], NISC Comparative Sequencing Program [b], Heidi H. Kong [c], Julia A. Segre [a†]

[a] Microbial Genomics Section, Translational and Functional Genomics Branch, NHGRI, NIH, Bethesda, Maryland, USA

[b] NIH Intramural Sequencing Center, NHGRI, NIH, Rockville, Maryland, USA

[c] Cutaneous Microbiome and Inflammation Section, NIAMS, NIH, Bethesda, Maryland, USA

[†] Corresponding author

Julia A. Segre

Building 49, Room 4a26,

49 Convent Dr.

Bethesda, MD 20892

(301) 402-2314

jsegre@nhgri.nih.gov

This PDF file includes:

- SI Materials and Methods
- Tables S1, S2, S3, S4, S5 and S6
- Legends for Dataset S1, Dataset S2, Dataset S3, Dataset S4, Dataset S5, Dataset S6
- SI References
- Figures S1 to S12

MATERIALS AND METHODS

Healthy Volunteer recruitment and sampling

Healthy adult male and female volunteers 18–40 years of age were recruited from the Washington, DC metropolitan region. This natural history study was approved by the Institutional Review Board of the National Human Genome Research Institute (clinicaltrials.gov/NCT00605878) and the National Institute of Arthritis and Musculoskeletal and Skin Diseases (clinicaltrials.gov/NCT02471352) and all volunteers provided written informed consent prior to participation. Sampling was performed as described previously (1).

16S rRNA amplicon sequencing and analysis

Sample processing and sequencing were performed as described previously (1). Sequencing data was curated to include samples from fourteen distinct body sites collected from twenty-two healthy volunteers (HVs) (**Fig. S1**). 16S rRNA amplicon (V1-V3) sequencing data were processed using the DADA2 pipeline version v1.2.0 (2) to produce abundance estimates for error corrected amplicon sequence variants (ASVs). Reads were truncated to 375 nt with a max expected error (maxEE) of 2. The dada command (DADA2) was used to infer the composition of the sample with additional parameters specific to pyrosequencing data (HOMOPOLYMER_GAP_PENALTY=-1, BAND_SIZE=32). Chimeras were removed with the removeBimeraDenovo command. Taxonomic assignments were made using the DADA2 assignTaxonomy command with a curated RefSeq database (https://zenodo.org/record/3266798) and minimum bootstrap of 75. Taxonomic assignments were refined using BLAST to correct cases where a given ASV doesn't uniquely identify a species (see below). While we have shown previously that human-associated staphylococci can be differentiated using V1-V3 sequence data of sufficient length (3), we nevertheless tested the accuracy of our species-level assignments using simulations. Briefly, 74 Type sequence *Staphylococcus* isolates from the Ribosomal Database Project (4) were used to generate simulated sequences with an error profile similar to the one observed for pyrosequencing: 0.5% base changes, 0.1% inserts, 0.1% deletions (n=10 per reference sequence). Simulated sequences were trimmed to a 375 base V1-V3 region analogous to the real data and classified. The following species could not be separated, due to lack of variation in the selected V1-V3 region: *S. capitis/S. caprae*, *S. agnetis/S.* hyicus, *S. pseudointermedius/S. intermedius/S. delphini*. In addition, *S. argenteus* isn't distinguishable from *S. aureus* using V1-V3; however,

these two species are in general known to be difficult to differentiate (5). In addition to these, we could not assign taxonomy to two prominent species numbered Species-4 and Species-45. These were labelled as *S. saccharolyticus* group and *S. hemolyticus* group, respectively, denoting their closest taxonomically related species. ASV abundance estimates were then merged with sample metadata into a phyloseq (6) object for further processing.

Statistical analysis of 16S rRNA amplicon sequencing data

All analyses were performed in RStudio using the R v4.2.0 on the phyloseq object. For staphylococcal community analyses, the phyloseq object was trimmed to retain *Staphylococcus* ASVs using the subset_taxa function. ASVs that were present at ≥ 1% relative abundance in at least one sample were included. Dataset was further filtered to only retain ASVs present in more than three volunteers. This resulted in a final phyloseq object with 71 unique ASVs present in 298 samples. Read count for these 71 ASVs ranged from 99 to 176996. ASVs that belonged to the same species were merged using the tax_glom function within phyloseq, resulting in a total of 17 staphylococcal species. The distribution of these species was displayed using barplots, prevalence-abundance dotplots and boxplots. Non-parametric Wilcoxon rank-sum test was used to calculate statistical differences between microbial populations at different body-sites. P values below a standard alpha value (P value < 0.05) were considered significant. For beta-diversity measurement, phyloseq object with unmerged ASVs (before tax-glom) was used. Arcsine transformation was used for variance stabilization of OTU counts. Beta-diversity was visualized by Principal Coordinate Analysis (PCoA) of the Bray-Curtis dissimilarity metric generated using vegdist in the vegan v2.6-2 package. Effect size of individual ASVs was calculated with the envfit function (vegan). Variation by body-site was measured using PERMANOVA with the adonis function (vegan), using 10,000 permutations.

Isolate collection

Skin and nasal cultures were obtained with Catch-all Collection Swabs (Epicentre) pre-moistened with Fastidious Broth (Remel), placed in 2.0ml Fastidious Broth supplemented with 10% glycerol, and frozen at -80°C. Swabs were thawed, vortexed, serial diluted, and plated on Tryptic Soy Agar with 5% Sheep Blood (Remel). After overnight incubation at 37°C, colonies were picked and stored in TSB with 20% glycerol. Colonies were screened by PCR for *S. capitis* using ScapF (5'- GCTAATTTAGATAGCGTACCTTCA -3') and ScapR (5'- CAGATCCAAAGCGTGCA -3') (7), *S. epidermidis* using Se705-1 (5'- ATCAAAAAGTTGGCGAACCTTTTCA -3') and Se705-2 (5'-

CAAAAGAGCGTGGAGAAAAGTATCA -3') (7) and *S. hominis* using hom-F (5'-TACAGGGCCATTTAAAGACG - 3') and hom-R (5'- GTTTCTGGTGTATCAACACC -3") (8). Species taxonomy of isolates was confirmed by sequencing 16S rRNA gene using Sanger sequencing.

Whole genome sequencing

Individual staphylococcal colonies were streaked on blood agar for two passages. Isolates were grown overnight in Tryptic Soy Broth at 37° C, pelleted with centrifugation, and genomic DNA was extracted using the Promega Maxwell Tissue DNA Kit with the addition of Readylyse Lysozyme Solution (Epicentre) and Lysostaphin (Sigma). DNA was treated with RNase, re-purified with the Genomic DNA Clean and Concentrator Kit (Zymo), and quantified using a Nanodrop spectrophotometer and Qubit (ThermoFisher). 1.0ng of bacterial DNA was used as input into the Nextera XT Sample prep kit (Illumina) as suggested by manufacturer. Nextera libraries were generated from the genomic DNA and sequenced using a paired-end 300-base dual index run on an Illumina MiSeq to generate 1 million to 2 million read pairs per library for ~80x genome coverage. 79 *S. capitis*, 73 *S. epidermidis*, and 121 *S. hominis* isolates (273 total) from eighteen body sites of fourteen healthy volunteers were sequenced.

Whole genome assembly

Nextera libraries for each isolate were multiplexed on a Novaseq 6000. Reads were subsampled to 80x coverage using seqtk (version 1.2), assembled with SPAdes (version 3.14.1) (9) and polished using bowtie2 (version 2.2.6) and Pilon (version 1.23) (10). To achieve full reference genomes for select isolates, genomic DNA was sequenced on the PacBio Sequel II platform (version 8M SMRTCells, Sequel II version 2.0 sequencing reagents, 15 hr movie collection). The subreads were assembled using Canu v2.1 (11) and polished using the pb_resequencing workflow within PacBio SMRTLink v.9.0.0.92188. Oxford Nanopore Technologies (ONT) sequence reads were randomly downsampled to an estimated 200x coverage using seqtk version 1.2 and assembled using canu-2.2 using the -fast parameter. The draft contigs were polished using Racon version 1.5.0 (12) from a sorted alignment of the ONT reads generated using minimap2 version 2.24 (13) and samtools version 1.11. The Racon-polished contigs were circularized and/or joined by manually evaluating contig overlaps using Gepard v1.30 (14). The circularized chromosome and plasmid sequences were polished with Medaka version 1.6.1 using the sampled ONT sequence reads. The sequences were polished further with Pilon version 1.23 (10) from a sorted alignment generated with BWA-MEM version

0.7.17 and samtools using Illumina paired-end reads trimmed with Trimmomatic version 0.39 (15). Genome annotation was performed using National Center for Biotechnology Information (NCBI) Prokaryotic Genome Annotation Pipeline (PGAP: https://www.ncbi.nlm.nih.gov/genome/annotation_prok/).

Genome quality control

Genomes with ≥ 95% fastANI distances were considered to belong to the same species (16). All 273 genomes were quality filtered according to completeness (≥ 98%) and contamination (≤ 5%) using CheckM (17). dRep v3.2.2 (18) was used to de-replicate and collapse highly similar genomes using a >99.9% ANImf threshold and a minimum alignment coverage of 50%, as used previously (19). A representative genome from each clonal cluster was chosen for pan-genome analysis based on dRep's score-based system, which incorporates completeness, contamination, strain heterogeneity, N50, genome length and centrality. The following command was used: *dRep dereplicate -p 10 -comp 98 -con 5 -sa 0.999 -nc 0.5 -g *.fasta*

Species-level pan-genome

Dereplicated draft genomes, including plasmids, were annotated using prokka v1.14.6 (20). A customized protein database was generated for prokka annotation using the nine staphylococcal reference genomes sequenced in this study using either the PacBio or Oxford Nanopore technology. This was to ensure consistent annotation across all genomes. Prokka generated *gff3 output was passed on to panaroo v3.1.2 (21) for species-level pan-genome analysis. Panaroo was run under the strict mode without merging paralogs. The initial sequence identity threshold was 98% with subsequent clustering into protein families being performed using a threshold of 70% identity. Core genes were defined using a 99% presence threshold. A core gene alignment was generated using the default mafft aligner and the -a flag, which was used for generating phylogenetic trees. Panaroo command: *panaroo -i *.gff -o results -clean-mode strict -refind_prop_match 0.7 -alignment core -codons -t 24*

Genus-level pan-genome

We employed a reciprocal-best-blast approach to cluster the species-level pan-genome references (containing a representative sequence from each orthologous gene cluster detected in the pan-genome) at the protein level (≥ 40% identity and ≥ 50% coverage). A > 50% breadth of coverage cutoff was chosen to minimize domain-level alignments. The percent identity threshold of > 40% (amino acid) was chosen by looking at the genus core size as a function of

the identity threshold (90% down to 30%). We observed an increase in the size of the core down to 40% identity. This 40% cutoff in the context of a reciprocal best hit strategy where orthologs from all three species needed to be included, maximized clustering of expected orthologs based on annotated gene names and functions. This resulted in clustering of 78-93% of core genes from each species-level pan-genome to provide a merged genus pan-genome**.**

Pan-genome analyses

Gene presence absence matrices from Panaroo and three-way-reciprocal blast were used for individual species and genus pan-genome analyses, respectively, using R v4.2.0. micropan package v2.1 (22) was run with default parameters for principal component analysis (PCA). Cluster analysis to identify genome grouping based on gene content was performed using dist() and hclust functions in base R. pheatmap v1.0.12 was used for generating heatmaps. Venneuler 1.1.3 was used to make proportional Venn diagrams. Tanglegram showing phylogeny versus pan-genome tree was generated using dendextend package v1.16.

Functional annotation and reconstruction of metabolic pathways

Prokka provided default annotations for most genes. In addition to this, thorough functional annotations and pfam domain predictions were carried out by eggNOG-mapper v2 (web version) using eggNOG 5.0 database, DIAMOND aligner, and the following flags "–evalue 0.001 –score 60 –pident 40 –query_cover 20 –subject_cover 20 –itype CDS –translate –tax_scope auto –target_orthologs all –go_evidence all –pfam_realign denovo". Panaroo output file pan_genome_reference.fa, containing reference sequences of all the genes found in a species pan-genome, was used as an input for annotation.

Completeness of Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways was tested using a previously described pipeline (23). Bacterial pathways that showed $\geq$ 75% completeness in at least one genome were kept for downstream analyses.

Functional enrichment analysis of KEGG pathways was performed using MicrobiomeProfiler, a R-shiny package based on clusterProfiler (24). KO identifiers assigned by eggNOG-mapper v2 were used as input for the analysis.

Phylogenetic tree construction

Phylogenetic tree for each species was constructed using the core-gene alignment generated by panaroo. ModelFinder (25) implemented in IQ-TREE 2 (26) was employed to find the best substitution and site heterogeneity models. GTR+F+I+G4, GTR+F+R2 and GTR+F+R5 models

were selected for *S. capitis*, *S. epidermidis*, and *S. hominis*, respectively to construct maximum likelihood (ML) trees using RAxML-NG v1.1.0 (27) with 1,000 bootstrap replicates. Each species tree was rooted using genomes from the other two species as an outgroup. For this rooting, a genus-level phylogenetic tree was built using GTR+F+R5 model on the core-gene alignment of 665 genus-core genes generated by panaroo using all 126 genomes and the same threshold values as outlined above. iTOL v6 (28) was used for tree display and annotation.
RAxML-NG command: *raxml-ng -all -msa core_gene_alignment_filtered.aln -model -prefix -seed -threads -bs-metric fbp,tbe -bs-trees 1000*

pan-genome-wide association study

Scoary v1.6.16 (29) was used to identify overrepresentation of genes within each phylogenetic clade for individual species. The input trait file consisted of dummy variables (0 and 1) to denote clade A and clade B membership. gene_presence_absence_roary.csv output from panaroo was used for gene content differences between species genomes. Scoary was run for each species using the following command: *scoary -g gene_presence_absence_roary.csv -t traits_clades.csv -no_pairwise -o results p 0.05 -c BH --threads 10*

Isolate selection for growth curves:

A greedy algorithm was used to choose a subset of staphylococcal isolates for growth curve analysis. Briefly, *gene_presence_absence.Rtab* output file from panaroo was used to iteratively select genomes that best improve the representation of the species-level pan-genome. Initially, genomes were selected to cover the core+softcore+shell genes (genes found in at least 15% of genomes) using the minimum number of genomes. Then additional isolate genomes were added to increase the coverage of cloud genes (present in <15% of genomes) such that > 88% of the pan-genome (without singletons) was encoded by final isolate genomes. In some cases, we forced the inclusion of genomes that were phenotypically valuable or had finished genomes. The final dataset included 15 *S. epidermidis* isolates (75.4% total pan-genome; 96.1% without singletons), 12 *S. capitis* (94.1%; 99.6%), and 14 *S. hominis* isolates (68.2%; 88.7%).

Growth analysis:

All isolates were cultivated overnight, at 37 °C aerobically, in Brain Heart Infusion with 0.5% Yeast extract broth (BHI-YE) from frozen glycerol stocks. Overnight cultures were used as inocula (1:100) for BHI-YE and artificial skin media ES and ESL (Pickering Laboratories Inc Catalog #s 1700-0023 and 1700-0556 respectively), which were distributed in 96-microwell

Nunc™ Edge plate (Catalog # 267544) (200ul/well in duplicate). Plates were sealed using Parafilm™ to reduce evaporation and incubated in BioTek Epoch 2 microplate reader at 34° C with continuous shaking at 180 rpm. Kinetic growth data was monitored by OD measurement at 600 nm every 30 minutes for up to 24 hours. A minimum of 4 biological replicates were generated for each isolate-medium pair by repeating the growth curves on separate days. Biological replicate data was pooled and area under the curve (AUC) quantified using Growthcurver (30). Data analyses and visualization was carried out using the R software. Composition of ES and ESL media are given in **Table S5.**

RNA extraction and library preparation:

Isolates for RNA-Seq analysis were grown following the same protocol as for growth curve analysis, except for using an Eppendorf deep 96-wells plate (2 ml) to accommodate larger culture volumes necessary to harvest sufficient RNA for sequencing. Cultures were grown to mid-log phase, and up to 6 ml of culture per isolate per medium was harvested using Bacterial RNAprotect (Qiagen 76506). Three biological replicates were generated for each condition. For RNA extraction we used ZymoBiomcs DNA/RNA Miniprep Kit (R2002) to isolate bacterial RNA by adding 600μl RNA Shield to resuspend the bacterial pellets. After following all the steps outlined in the kit catalog, final total RNA was eluted in with 50μl nuclease free water. An extra DNase step was added to ensure complete removal of genomic DNA contamination. This involved digestion with 1μl of rDNase (Thermo Fisher AM2222) at 37°C for 30 min, followed by inactivation with 0.1 volume of DNase inactivation reagent. The extracted total RNA was cleaned up using NEB's Monarch RNA Cleaning kit (T2030L), then eluted with 10-20μl nuclease-free water. RNA was quantified by Nanodrop and Qubit and the quality check was performed using bioanalyzer. Ribosomal RNA was removed by Qiagen's FastSelect - 5S/16S/23S (335927). RNA fragmentation was set at 89°C for varying times depending on sample RIN (RNA integrity number), then gradually cooled using a gradient from 75°C to 25°C. After cleanup using QIAseq, Illumina stranded total RNA prep kit (20040529) was used to synthesize first and second strand cDNA. Agencourt AMPure XP beads (A63881) were used for cDNA cleanup. Adenine nucleotide was added to the 3' ends for ligating RNA-index anchors to the double-stranded cDNA fragments, followed by another cleanup using AMPureXP beads. Selective PCR amplification of the anchor-ligated DNA fragments was done using unique-dual Index (20627581) to generate an RNA-Seq library. Libraries were assayed for quality and concentration by Agilent 2100 Bioanalyzer and DNA 1000 kit, then sent to NISC for sequencing.

Bioinformatic analysis of RNA-seq data

Libraries were sequenced on the Illumina NovaSeq 6000 platform at the NIH Intramural Sequencing Center to a target depth of 50 million 2x150 paired-end reads per sample. Reads were trimmed for adapters with cutadapt v3.4 using the parameters "--nextseq-trim 20 -e 0.15 -m 50" (https://journal.embnet.org/index.php/embnetjournal/article/view/200) and checked for quality with PRINSEQ-lite v0.20.4 using the parameters "-lc_method entropy -lc_threshold 70 -min_len 50 -min_qual_mean 20 -ns_max_n 5 -min_gc 10 -max_gc 90" (31). Reads with less than 50 bp length after trimming were removed.

Rockhopper v 2.0.3 (32) was used to align reads from each sample to a reference genome. Each isolate had its own reference genome to ensure alignment of non-core gene reads. Raw read counts per gene were obtained from rockhopper, and those mapping to ncRNAs, rRNAs and tRNAs were removed from the analysis. Initial analysis was performed using a subset of 1647 genus-core genes that had at least one read in all samples. For this purpose raw reads were normalized using the estimateSizeFactors() function and transformed using the varianceStabilizingTransformation() function in the DESeq2 package (33) prior to analysis. Principal components were determined using the prcomp() function. DESeq2 was used to calculate differentially expressed genes ($\log_2$ fold change ≥ 1 or ≤ −1 and *adjusted P value* < 0.05) in ES medium relative to BHI-YE in individual isolates to measure the contribution of all the genes encoded within a genome. Differential expression was determined using 'apeglm' for LFC shrinkage (34). Heatmaps were generated using pheatmap and ComplexHeatmap R packages.

Supplementary Table 1. Prevalence and mean relative abundance of skin-resident staphylococcal species based on 16S rRNA amplicon analysis.

Species	Mean Percent Relative Abundance* ± Standard Deviation	Prevalence by subjects (n = 22) (%)	Prevalence by samples (n = 298) (%)
S. epidermidis	52.25 ± 1.87	22 (100)	284 (95.30)
S. capitis	26.37 ± 1.77	22 (100)	225 (75.50)
S. hominis	23.65 ± 1.82	22 (100)	208 (69.80)
S. warneri	8.27 ± 1.23	22 (100)	121 (40.60)
S. lugdunensis	3.98 ± 1.69	17 (77)	31 (10.40)
S. haemolyticus	6.27 ± 1.49	16 (73)	53 (17.79)
S. auricularis	30.79 ± 7.03	13 (59)	27 (9.06)
S. cohnii	12.19 ± 3.25	13 (59)	35 (11.74)
S. pettenkoferi	9.19 ± 1.92	9 (41)	19 (6.38)
S. aureus	8.44 ± 2.83	8 (36)	17 (5.70)
S. saccharolyticus	16.32 ± 3.47	6 (27)	31 (10.40)
S. saccharolyticus group	16.03 ± 2.88	6 (27)	27 (9.06)
S. saprophyticus	0.84 ± 0.28	6 (27)	8 (2.68)
S. simulans	3.31 ± 1.14	5 (23)	6 (2.01)
S. pasteuri	3.26 ± 1.06	5 (23)	14 (4.70)
S. haemolyticus group	4.48 ± 1.35	4 (18)	13 (4.36)
S. petrasi	0.86 ± 0.22	4 (18)	7 (2.35)

*Mean percent relative abundance of each species was calculated using only those samples that were positive for the given species (range: minimum 3 reads to maximum 12119 reads).

Supplementary Table 2. Genome characteristics of each species based on isolates used in this study.

Genome characteristic	***S. epidermidis*** **(N = 49)**	***S. capitis*** **(N = 22)**	***S. hominis*** **(N = 55)**
Mean Genome Size (Mb)	2.507317 ± 0.085426	2.480510 ± 0.049336	2.273876 ± 0.077851*
Mean GC-content (%)	36.7 ± 0.2	36.9 ± 0.3	36.2 ± 0.2
Mean number of protein coding genes (CDS)	2315 ± 88	2378 ± 47	2195 ± 93**

*, ** *S. hominis* has a significantly smaller genome and encodes fewer CDSs than the other two species surveyed ($p < 0.01$, one-way ANOVA, post hoc Tukey's HSD)

Supplementary Table 3. Major pan-genome features of each species

Pan-genome Feature	***S. epidermidis*** **(N = 49)**	***S. capitis*** **(N = 22)**	***S. hominis*** **(N = 55)**
Total Size (Mb)	3.67	2.71	3.91
Total genes	4699	3203	5118
Core genes (%)	1902 (40.47%)	2091 (65.28%)	1761 (34.40%)
Soft-core genes (%)	50 (1.06%)	12 (0.37%)	72 (1.40%)
Shell genes (%)	556 (11.83%)	477 (14.89%)	585 (11.43%)
Cloud genes (%) (Singletons)	2191 (46.63%) (987)	623 (19.45%) (290)	2700 (52.75%) (1265)
Mean percent core genes per genome	83.87 ± 2.69	89.79 ± 1.64	81.19 ± 3.16

Table 4. Genus-core genes with predicted role in skin colonization

Gene ID	Genes	Function	Role in skin colonization
S0510	*srtA*	covalent anchoring of adhesins to the bacterial cell wall	bacterial adhesion, biofilm formation, and immune escape (35)
S1463, S0384, S0385, S1644	*dltABCD*	D-alanylation of teichoic acids	Resistance to host antimicrobial peptides (36)
S0319	*mprF*	phospholipid lysylation	Resistance to host antimicrobial peptides (37)
S0957	*sepA*	protease	AMP degradation (38)
S1750, S1576	*vraF, vraG*	AMP export	Resistance to host antimicrobial peptides (39)
S1493, S1721, S1590	*graR, graS, graX*	Aps system	AMP sensor, regulator of AMP resistance mechanisms (40)
S0518- S0520	*capABC*	Poly-γ-DL-glutamate capsule biosynthesis	Protects from AMPs, phagocytosis, and high salt concentration (41)
S1387	*oatA*	O-acetylates peptidoglycan	Lysozyme resistance (42)
S0663	*vraX*	binds host complement protein	Inhibit classical complement pathway (43)
S1611	*atlE*	Bifunctional autolysin/adhesin	biofilm formation, vitronectin binding (44)
S1620, S1526, S1030, S1031	*pmtABCD*	ABC transporter for all Phenol-soluble modulins (PSM) classes	Virulence, biofilm (45)
S0455, S1812	-	Phenol-soluble modulins-beta	promote biofilm maturation and dissemination (46)

S0624, S0625, S1613, S1766, S0626, S0627, S1813	*tagDXBGHA*	poly-glycerol-phosphate teichoic acids synthesis	Role in nasal colonization (47)
S1337	*gehC*	lipase	establish residence in the hair follicles (48)
S0557	*betA*	Choline dehydrogenase	biosynthesis of the osmoprotectant glycine betaine (49)
S0487, S0486, S0485, S1232	*OpuCABCD*	Glycine betaine/choline/carnitine transport	Osmoprotolerance (50)
S1313, S0867-S0872 S0873	*ureABCEFGD* *yut*	urease-production urea transport	acid response and pH homeostasis (51)
S0947- S0949	*yfmCDE*	ferric citrate transporter	Iron acquisition (52)
S0951	*isdG*	liberates iron from host heme	Iron acquisition (53)

Supplementary Table 5. Composition of artificial skin media
(www.pickeringlabs.com/)

1. Eccrine Sweat (ES):

Artificial Eccrine Perspiration (pH 5.5); Catalog Number: 1700-0023

Amino Acids
Concentrations for listed amino acids range from 0.002 g/L (for Taurine) to 0.30 g/L (for Serine)

- Glycine
- L-Alanine
- L-Arginine
- L-Asparagine
- L-Aspartic acid
- L-Citrulline
- L-Glutamic acid
- L-Histidine
- L-Isoleucine
- L-Leucine
- L-Lysine as hydrochloride
- L-Methionine
- L-Ornithine as hydrochloride
- L-Phenylalanine
- L-Serine (Largest amount)
- L-Threonine
- L-Tyrosine
- L-Valine
- Taurine

Metabolites
Concentration for listed metabolites range from 0.015 g/L (for Uric Acid) to 1.74 g/L (for Urea)

- Uric acid
- Urea
- Lactic acid
- Ammonia

Minerals

- Sodium – 33 mmole/L
- Zinc – 11.21 µmole/L
- Chloride – 80.34 mmole/L
- Calcium – 5.49 mmole/L
- Iron – 4.62 µmole/L
- Magnesium – 1.67 mmole/L
- Potassium – 33 mmole/L
- Sulfate – 2.57 mmole/L

2. **Eccrine Sweat with lipids (ESL):** This medium was prepared using ES medium as the base. Tween 80 was added at 0.1%. The following sebum/apocrine emulsion was added at 1% for final growth curves.

   Apocrine sweat: emulsion of sebum plus other ingredients. Catalog Number 1700-0556

Compound	**Concentration (g/L)**
L-Alanine	0.1-0.5
L-Aspartic acid	0.01-0.1
L-Citrulline	0.1-0.5
L-Glutamic acid	0.5 -2
L-Glutamine	0.1-0.5
Glycine	0.1-0.5
L-isoleucine	0.1-0.5
L-Leucine	0.1-0.5
L-Lysine	0.5-2
L-Phenylalanine	0.01-0.1
L-Proline	0.1-0.5
L-Serine	0.1-0.5
L-Threonine	0.1-0.5
L-Tryptophan	0.1-0.5
L-Tyrosine	0.01-0.1
L-Valine	0.1-0.5
Creatine	0.01-0.1
Urea	0.2-2
Citric acid	0.1-0.5
Formic Acid	0.01-0.1
Lactic Acid	0.5 - 2
Glucose	0.01-0.1
Butyric acid	2-3
Valeric acid	2-3
a-hydroxy-n-butyric acid-sodium salt	0.01-0.1
3-hydroxybutyric acid	0.01-0.1
a-hydroxy-iso-butyric acid	0.1-0.5
$(NH_4)_2SO_4$	0.1-0.5
$CaCl_2*2H_2O$	0.5 - 2
$CuSO_4*5H_2O$	0.001-0.01
$Fe(NO_3)_3*9H_2O$	0.001-0.01
$MgCl_2*6H_2O$	0.1-0.5
NaCl	0.5 - 2
$ZnCl_2$	0.001-0.01

Sebum components

- Palmitic acid (CAS# 57-10-3): 0.5 % w/v
- Stearic acid (CAS# 57-11-4): 0.25% w/v
- Oleic Acid (CAS# 112-80-1): 0.9 % w/v
- Linoleic acid (CAS# 60-33-3): 0.25% w/v
- Coconut oil (CAS# 8001-31-8): 0.75 % w/v
- Olive oil (CAS# 800-25-0): 1% w/v
- Paraffin Wax (CAS# 8002-74-2): 0.5 % w/v
- Synthetic Spermacetti: 0.75% w/v
- Squalene (CAS# 111-02-4): 0.25% w/v
- Cholesterol (CAS# 57-88-5): 0.25% w/v
- Triethanolamine (CAS#102-71-6): 0.8% w/v

Supplementary Table 6. Distribution of differentially expressed genes in ES medium relative to BHI-YE between core and accessory gene partitions

Differentially expressed genes = Fold change ≥ or ≤ 2, adjusted P value < 0.05

Species	Isolate	No. of genus-core genes	Percent (%)	No. of species-restricted-core genes	Percent (%)	No. of accessory genes	Percent (%)
S. epidermidis	NIHLM087	583	72	149	18	81	10
	SENIH040	428	70	107	17	78	13
	SENIH047	398	64	119	19	104	17
S. capitis	SCNIH004	345	66	149	29	28	5
	SCNIH009	268	58	131	29	62	13
	SCNIH016	438	70	172	27	18	3
S. hominis	SHNIH025	371	70	70	13	92	17
	SHNIH027	166	74	32	14	26	12
	SHNIH030	463	70	85	13	115	17

Datasets in Excel format

Dataset S1. Genome statistics of isolates used in the current study

Dataset S2. Gene presence absence matrix based on individual species pan-genome

Dataset S3. Gene presence absence matrix of merged genus pan-genome

Dataset S4. List of species-restricted-core genes detected in genus pan-genome

Dataset S5. Pan-genome-wide-association study analysis showing clade-specific gene enrichment in each species

Dataset S6. List of differentially expressed genes in the ES medium relative to BHI-YE per isolate

REFERENCES

1. K. Findley *et al.*, Topographic diversity of fungal and bacterial communities in human skin. *Nature* **498**, 367-370 (2013).
2. B. J. Callahan *et al.*, DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods* **13**, 581-583 (2016).
3. S. Conlan, H. H. Kong, J. A. Segre, Species-level analysis of DNA sequence data from the NIH Human Microbiome Project. *PLoS One* **7**, e47075 (2012).
4. Q. Wang, G. M. Garrity, J. M. Tiedje, J. R. Cole, Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* **73**, 5261-5267 (2007).
5. R. Kaden, L. Engstrand, H. Rautelin, C. Johansson, Which methods are appropriate for the detection of Staphylococcus argenteus and is it worthwhile to distinguish S. argenteus from S. aureus? *Infect Drug Resist* **11**, 2335-2344 (2018).
6. P. J. McMurdie, S. Holmes, phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One* **8**, e61217 (2013).
7. T. Iwase, K. Seki, H. Shinji, Y. Mizunoe, S. Masuda, Development of a real-time PCR assay for the detection and identification of Staphylococcus capitis, Staphylococcus haemolyticus and Staphylococcus warneri. *J Med Microbiol* **56**, 1346-1349 (2007).
8. S. Hirotaki, T. Sasaki, K. Kuwahara-Arai, K. Hiramatsu, Rapid and accurate identification of human-associated staphylococci by use of multiplex PCR. *J Clin Microbiol* **49**, 3627-3631 (2011).
9. S. Nurk *et al.*, Assembling single-cell genomes and mini-metagenomes from chimeric MDA products. *J Comput Biol* **20**, 714-737 (2013).
10. B. J. Walker *et al.*, Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963 (2014).
11. S. Koren *et al.*, Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res* **27**, 722-736 (2017).
12. R. Vaser, I. Sovic, N. Nagarajan, M. Sikic, Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res* **27**, 737-746 (2017).
13. H. Li, Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100 (2018).
14. J. Krumsiek, R. Arnold, T. Rattei, Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics* **23**, 1026-1028 (2007).
15. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).
16. C. Jain, R. L. Rodriguez, A. M. Phillippy, K. T. Konstantinidis, S. Aluru, High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun* **9**, 5114 (2018).
17. D. H. Parks, M. Imelfort, C. T. Skennerton, P. Hugenholtz, G. W. Tyson, CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* **25**, 1043-1055 (2015).
18. M. R. Olm, C. T. Brown, B. Brooks, J. F. Banfield, dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J* **11**, 2864-2868 (2017).
19. M. R. Olm *et al.*, Identical bacterial populations colonize premature infant gut, skin, and

oral microbiomes and exhibit different in situ growth rates. *Genome Res* **27**, 601-612 (2017).
20. T. Seemann, Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068-2069 (2014).
21. G. Tonkin-Hill *et al.*, Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol* **21**, 180 (2020).
22. L. Snipen, K. H. Liland, micropan: an R-package for microbial pan-genomics. *BMC Bioinformatics* **16**, 79 (2015).
23. S. Saheb Kashaf *et al.*, Integrating cultivation and metagenomics for a multi-kingdom view of skin microbiome diversity and functions. *Nat Microbiol* **7**, 169-179 (2022).
24. T. Wu *et al.*, clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation (Camb)* **2**, 100141 (2021).
25. S. Kalyaanamoorthy, B. Q. Minh, T. K. F. Wong, A. von Haeseler, L. S. Jermiin, ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* **14**, 587-589 (2017).
26. B. Q. Minh *et al.*, IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol* **37**, 1530-1534 (2020).
27. A. M. Kozlov, D. Darriba, T. Flouri, B. Morel, A. Stamatakis, RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* **35**, 4453-4455 (2019).
28. I. Letunic, P. Bork, Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res* **49**, W293-W296 (2021).
29. O. Brynildsrud, J. Bohlin, L. Scheffer, V. Eldholm, Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biol* **17**, 238 (2016).
30. K. Sprouffske, A. Wagner, Growthcurver: an R package for obtaining interpretable metrics from microbial growth curves. *BMC Bioinformatics* **17**, 172 (2016).
31. R. Schmieder, R. Edwards, Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**, 863-864 (2011).
32. R. McClure *et al.*, Computational analysis of bacterial RNA-Seq data. *Nucleic Acids Res* **41**, e140 (2013).
33. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014).
34. A. Zhu, J. G. Ibrahim, M. I. Love, Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. *Bioinformatics* **35**, 2084-2092 (2019).
35. S. Cascioferro, M. Totsika, D. Schillaci, Sortase A: an ideal target for anti-virulence drug development. *Microb Pathog* **77**, 105-112 (2014).
36. A. Peschel *et al.*, Inactivation of the dlt operon in Staphylococcus aureus confers sensitivity to defensins, protegrins, and other antimicrobial peptides. *J Biol Chem* **274**, 8405-8410 (1999).
37. A. Peschel *et al.*, Staphylococcus aureus resistance to human defensins and evasion of neutrophil killing via the novel virulence factor MprF is based on modification of membrane lipids with l-lysine. *J Exp Med* **193**, 1067-1076 (2001).
38. Y. Lai *et al.*, The human anionic antimicrobial peptide dermcidin induces proteolytic defence mechanisms in staphylococci. *Mol Microbiol* **63**, 497-506 (2007).
39. M. Falord, G. Karimova, A. Hiron, T. Msadek, GraXSR proteins interact with the VraFG ABC transporter to form a five-component system required for cationic antimicrobial peptide sensing and resistance in Staphylococcus aureus. *Antimicrob Agents Chemother* **56**, 1047-1058 (2012).
40. M. Li *et al.*, The antimicrobial peptide-sensing system aps of Staphylococcus aureus. *Mol Microbiol* **66**, 1136-1147 (2007).

41. S. Kocianova *et al.*, Key role of poly-gamma-DL-glutamic acid in immune evasion and virulence of Staphylococcus epidermidis. *J Clin Invest* **115**, 688-694 (2005).
42. A. Bera, S. Herbert, A. Jakob, W. Vollmer, F. Gotz, Why are pathogenic staphylococci so lysozyme resistant? The peptidoglycan O-acetyltransferase OatA is the major determinant for lysozyme resistance of Staphylococcus aureus. *Mol Microbiol* **55**, 778-787 (2005).
43. J. Yan *et al.*, Staphylococcus aureus VraX specifically inhibits the classical pathway of complement by binding to C1q. *Mol Immunol* **88**, 38-44 (2017).
44. C. Heilmann, M. Hussain, G. Peters, F. Gotz, Evidence for autolysin-mediated primary attachment of Staphylococcus epidermidis to a polystyrene surface. *Mol Microbiol* **24**, 1013-1024 (1997).
45. S. S. Chatterjee *et al.*, Essential Staphylococcus aureus toxin export system. *Nat Med* **19**, 364-367 (2013).
46. R. Wang *et al.*, Staphylococcus epidermidis surfactant peptides promote biofilm maturation and dissemination of biofilm-associated infection in mice. *J Clin Invest* **121**, 238-248 (2011).
47. V. Winstel, P. Sanchez-Carballo, O. Holst, G. Xia, A. Peschel, Biosynthesis of the unique wall teichoic acid of Staphylococcus aureus lineage ST395. *mBio* **5**, e00869 (2014).
48. K. Nakamura, M. R. Williams, J. M. Kwiecinski, A. R. Horswill, R. L. Gallo, Staphylococcus aureus Enters Hair Follicles Using Triacylglycerol Lipases Preserved through the Genus Staphylococcus. *J Invest Dermatol* **141**, 2094-2097 (2021).
49. J. E. Graham, B. J. Wilkinson, Staphylococcus aureus osmoregulation: roles for choline, glycine betaine, proline, and taurine. *J Bacteriol* **174**, 2711-2716 (1992).
50. D. Casey, R. D. Sleator, A genomic analysis of osmotolerance in Staphylococcus aureus. *Gene* **767**, 145268 (2021).
51. C. Zhou *et al.*, Urease is an essential component of the acid response network of Staphylococcus aureus and is required for a persistent murine kidney infection. *PLoS Pathog* **15**, e1007538 (2019).
52. V. Braun, C. Herrmann, Docking of the periplasmic FecB binding protein to the FecCD transmembrane proteins in the ferric citrate transport system of Escherichia coli. *J Bacteriol* **189**, 6913-6918 (2007).
53. K. P. Haley, E. M. Janson, S. Heilbronner, T. J. Foster, E. P. Skaar, Staphylococcus lugdunensis IsdG liberates iron from host heme. *J Bacteriol* **193**, 4749-4757 (2011).

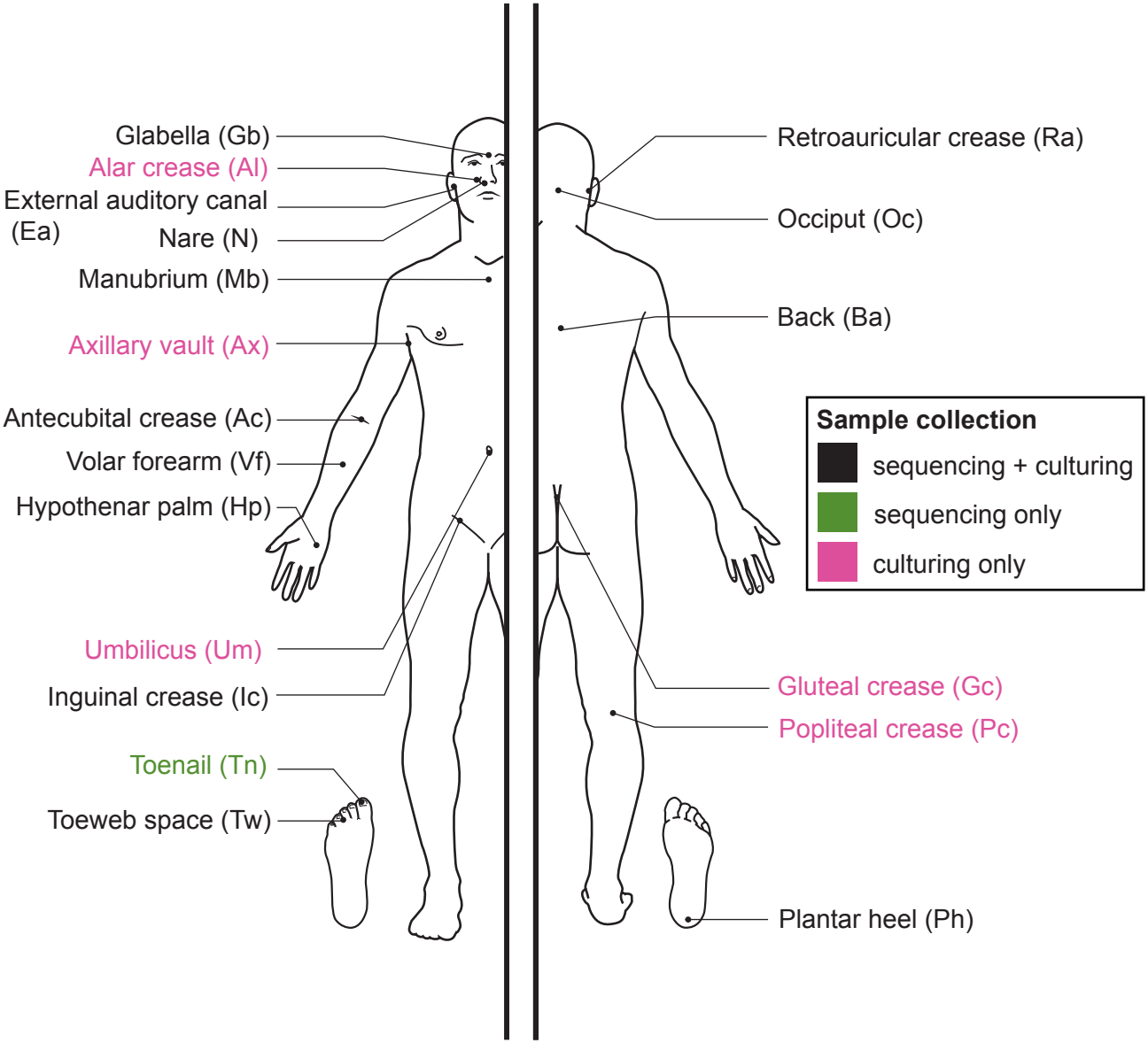

Figure S1. **Body sites used for collecting samples from healthy volunteers for 16S rRNA amplicon sequencing and culturing of staphylococcal isolates**.
Sites shown in black represent those for which both types of samples were collected. Toenail samples (green) were only collected for sequencing. Alar crease, Umbilicus, Gluteal crease, and Popliteal crease samples (all pink) were collected for culturing alone. Abbreviation shown in a bracket next to each body site were used to denote the site throughout the manuscript.

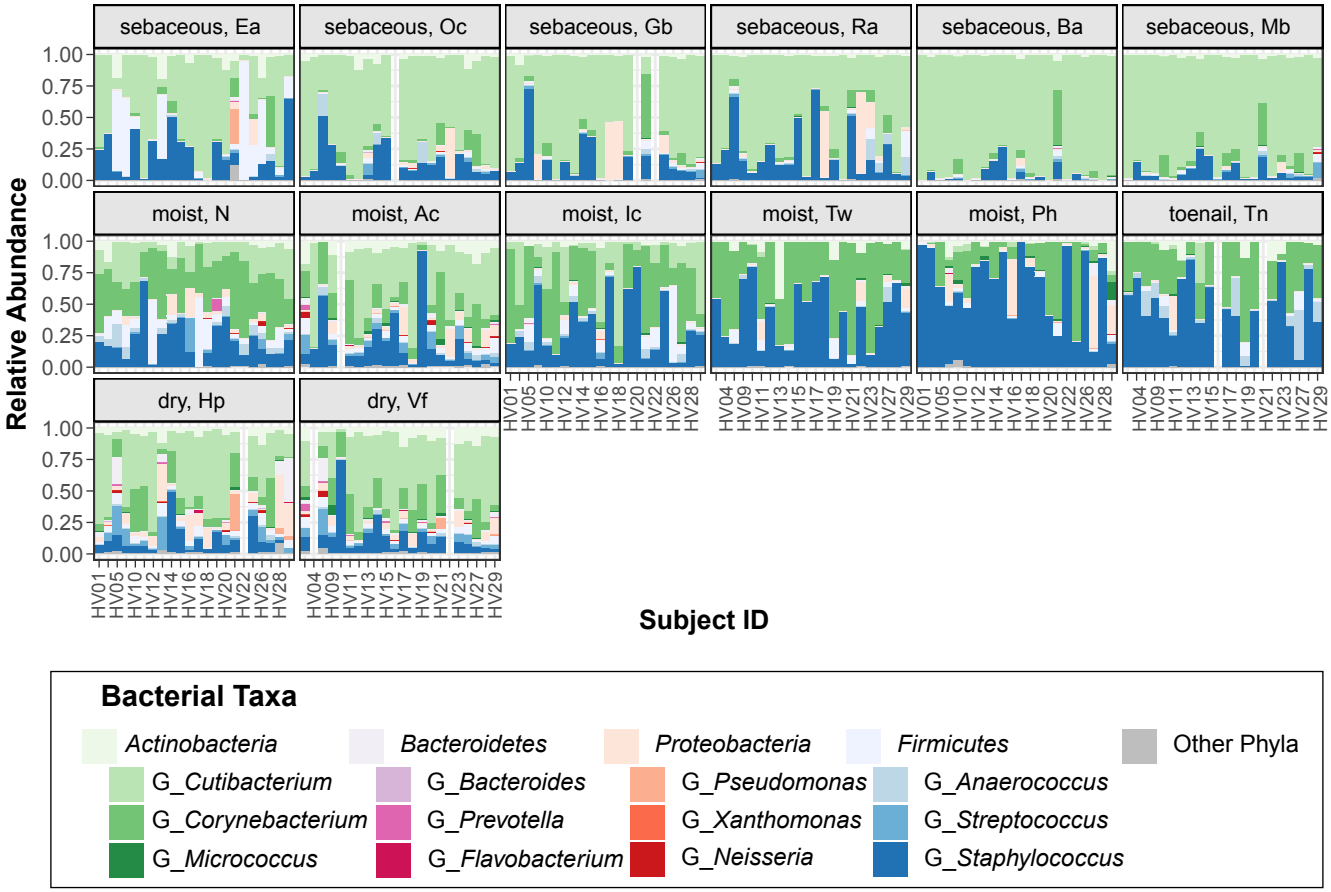

Figure S2. **Bacterial diversity on healthy human skin represented by major phyla and genera colonizing distinct body sites.**

Barplots display the relative abundance of major bacterial taxa at various body sites as displayed by facets. Colors represent taxa as shown in the accompanying legend. Each bar represents one volunteer. Empty bars represent missing data. Refer to Fig. S1 for body site abbreviation.

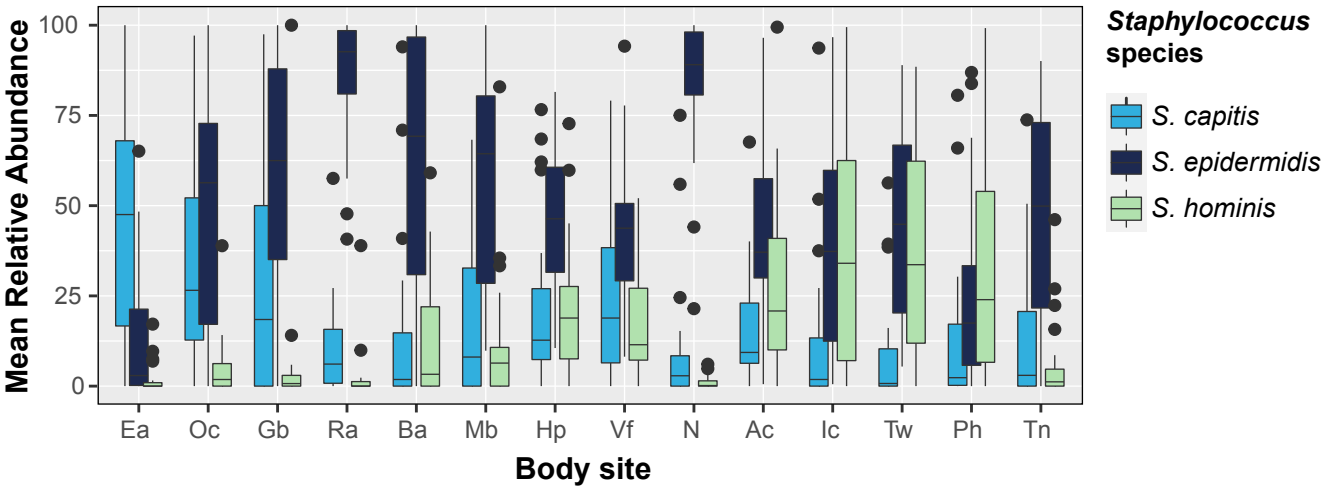

Figure S3. Mean relative abundance (MRA) of the three most prominent species within staphylococcal communities

Boxplots represents the MRA of the three staphylococcal species across all body sites as shown on the x-axis. Boxplot colors represent individual species. The center black line within each boxplot represents the median value, with edges showing the first and third quartiles. Refer to Fig. S1 for body site abbreviation.

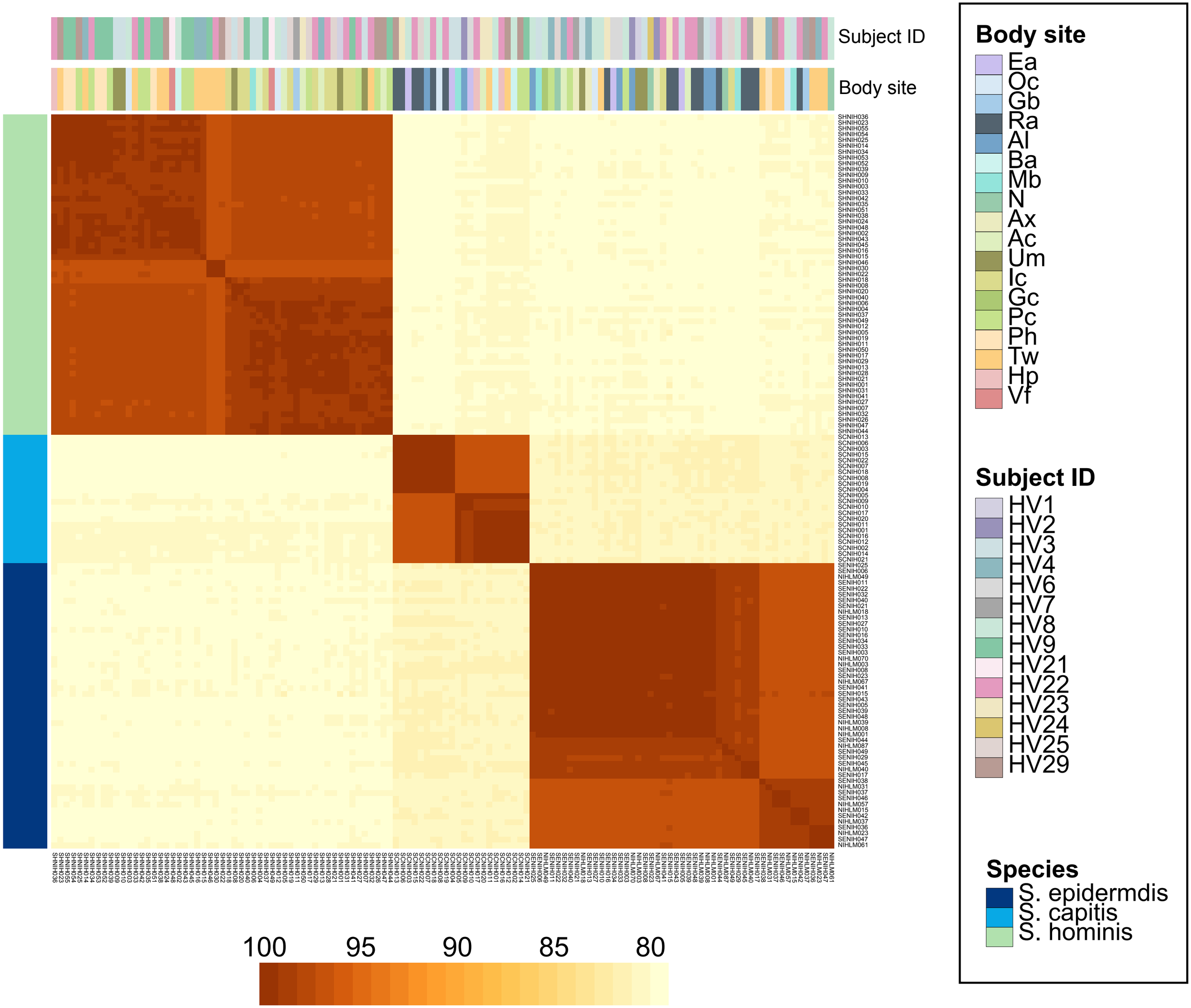

Figure S4. Heat-map of pairwise fastANI comparison of de-replicated, non-clonal genomes of *S. epidermidis*, *S. capitis*, and *S. hominis*.
Colored bars on the top denote the healthy volunteer and body site of isolation of each labeled genome. Species are denoted by the side bar on the left. Refer to Fig. S1 for body site details.

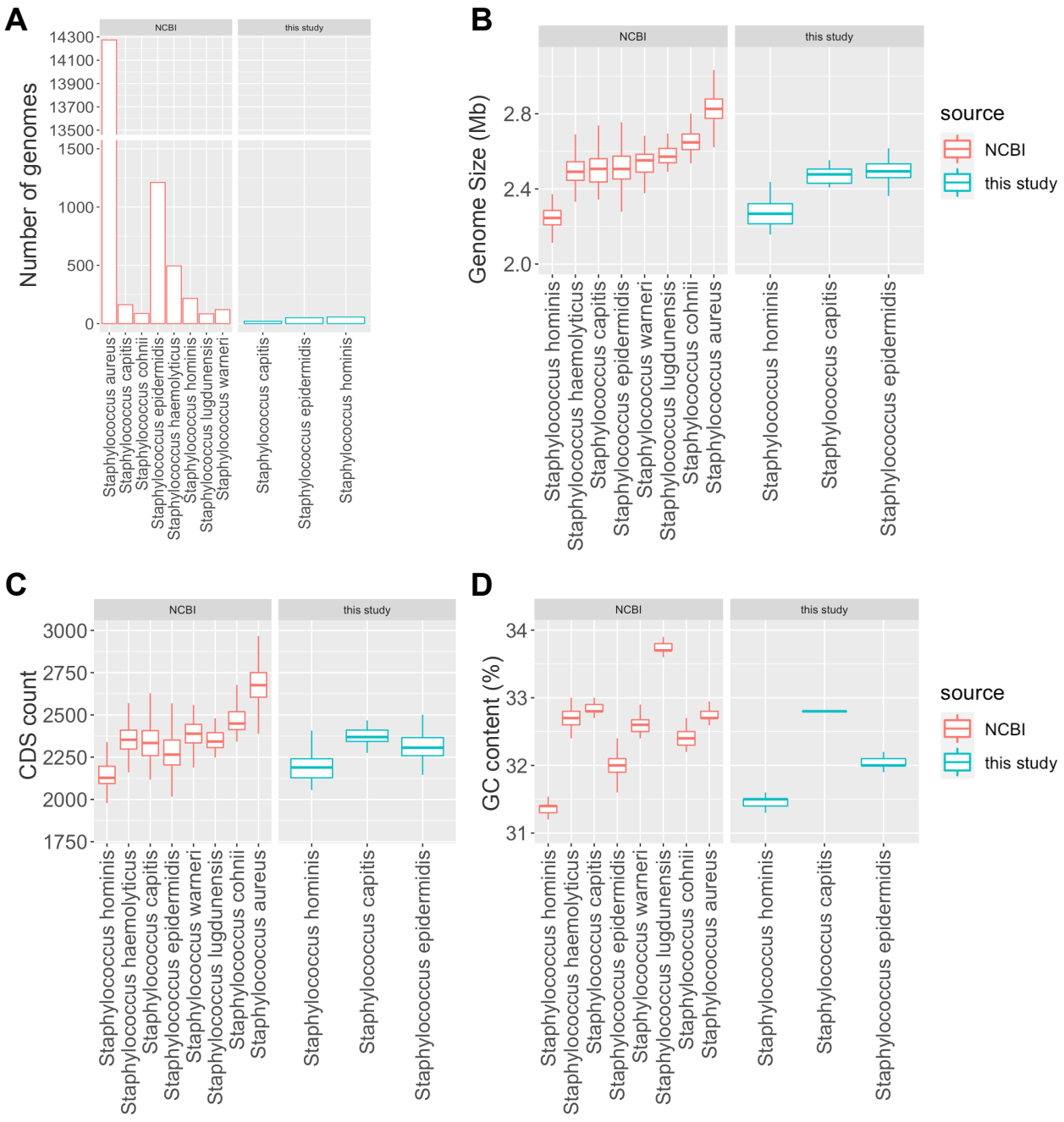

Figure S5. **Boxplot representing genome characteristics of isolates used in the current study and comparing them to species-specific genomes present in NCBI.**
(A) Number of genomes of different species that were either present in the NCBI or were sequenced for this study. **(B)** Average genome sizes of each species. **(C)** Average number of protein-coding genes (CDS) that were present in the genome of each species. **(D)** Percent GC content of each species.
The center line within each boxplot represents the median value, with edges showing the first and third quartiles.

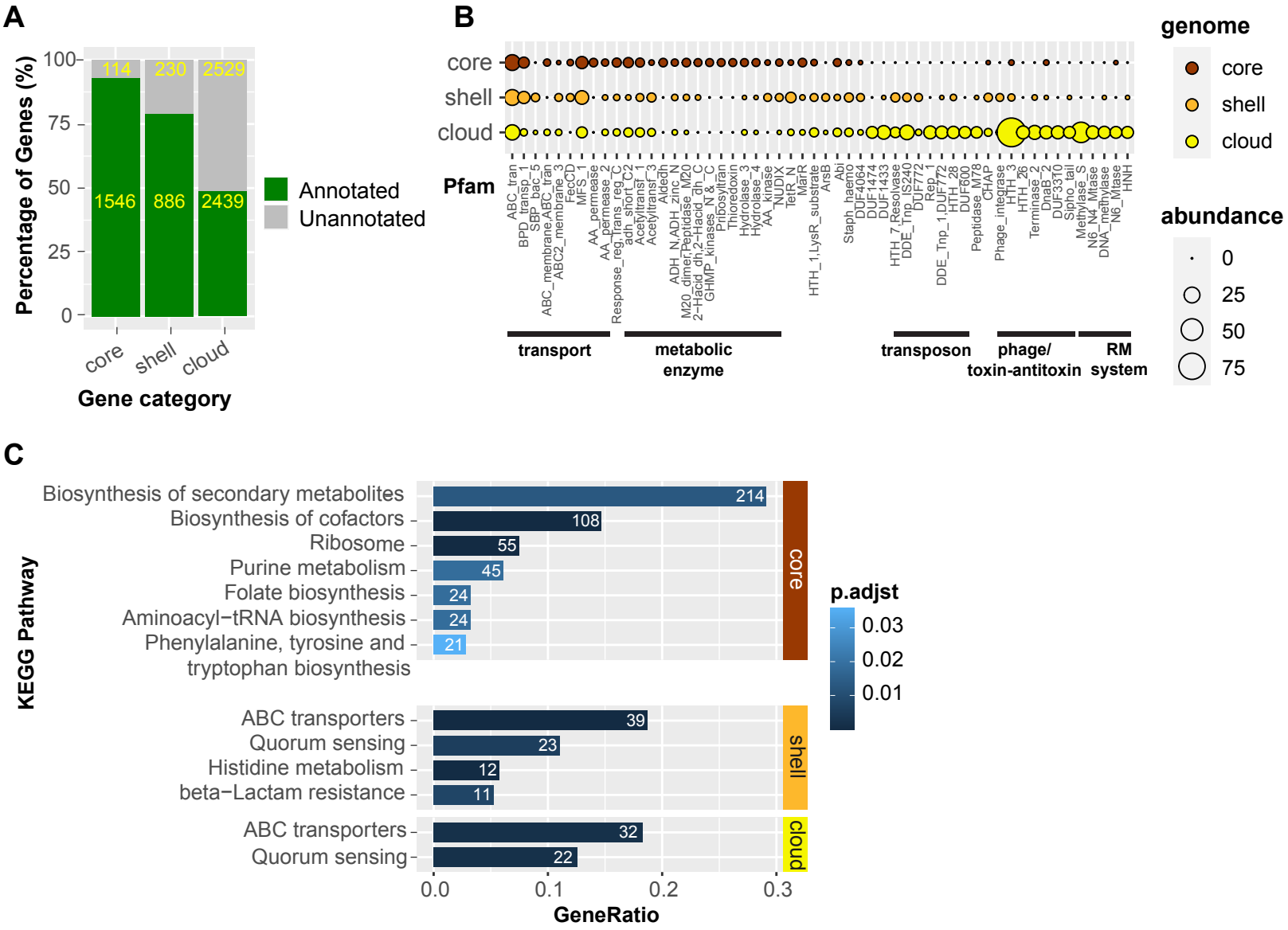

Figure S6. **Functional annotation of genus pan-genome.**

(A) Percentage of annotated genes shown by genus pan-genomic category. Actual number of genes in each group are shown in yellow inside each bar. **(B)** Combined data showing top 20 most represented Pfam domains in each pan-genomic category. Size of the bubbles represents the actual number of genes carrying the Pfam domain annotation within each category. **(C)** KEGG pathway enrichment analysis of genes in each category using KEGG KO identifiers.

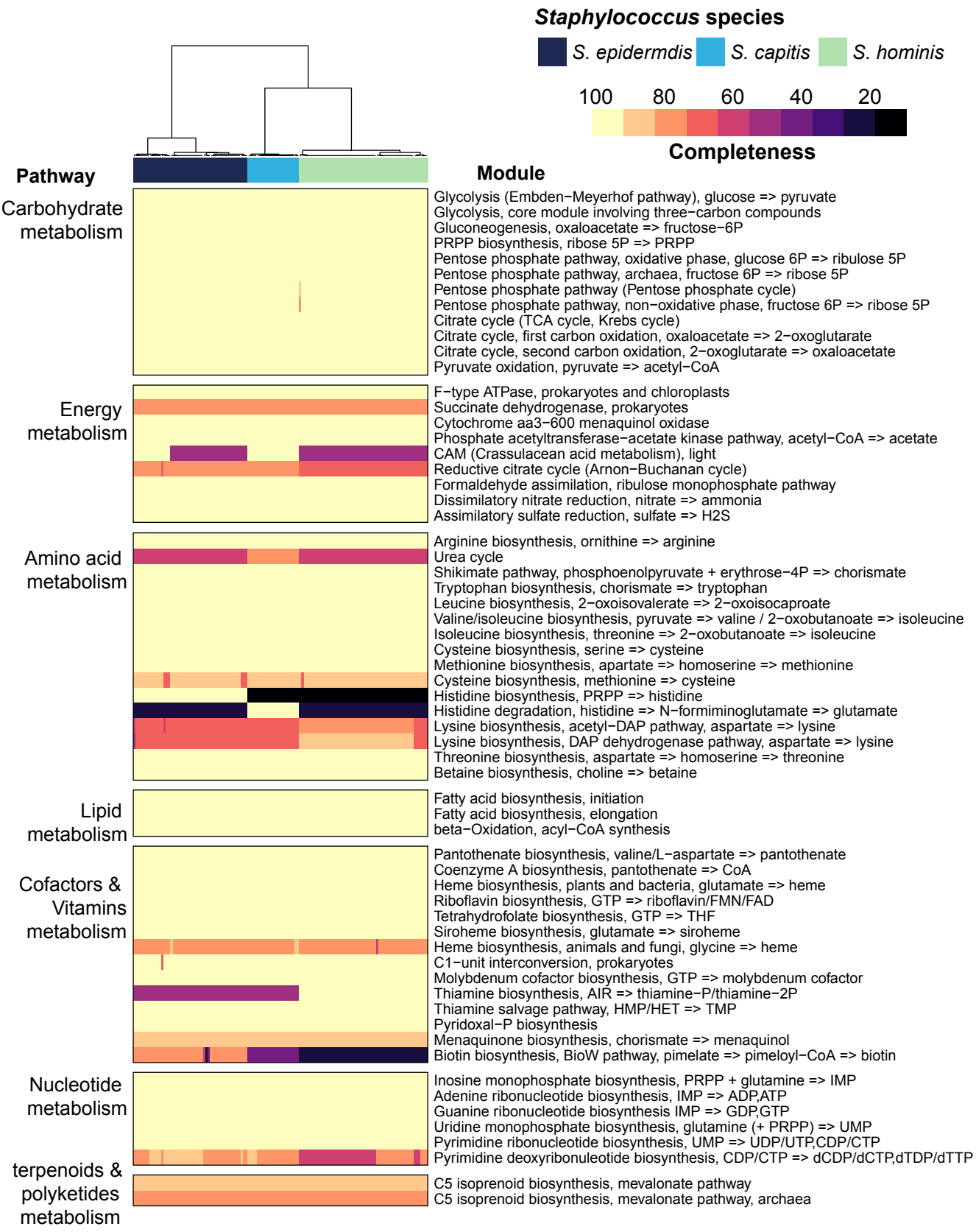

Figure S7. Core metabolic pathways present in staphylococcal genomes using KEGG module analysis.

Completeness of a pathway is expressed as a percentage from 0 to 100. Only pathways that showed ≥ 75% module completeness in at least one genome are shown. Hierarchical clustering (Euclidean distance; Ward) of species is shown as a dendrogram. Details of modules can be found at: https://www.genome.jp/brite/ko00002

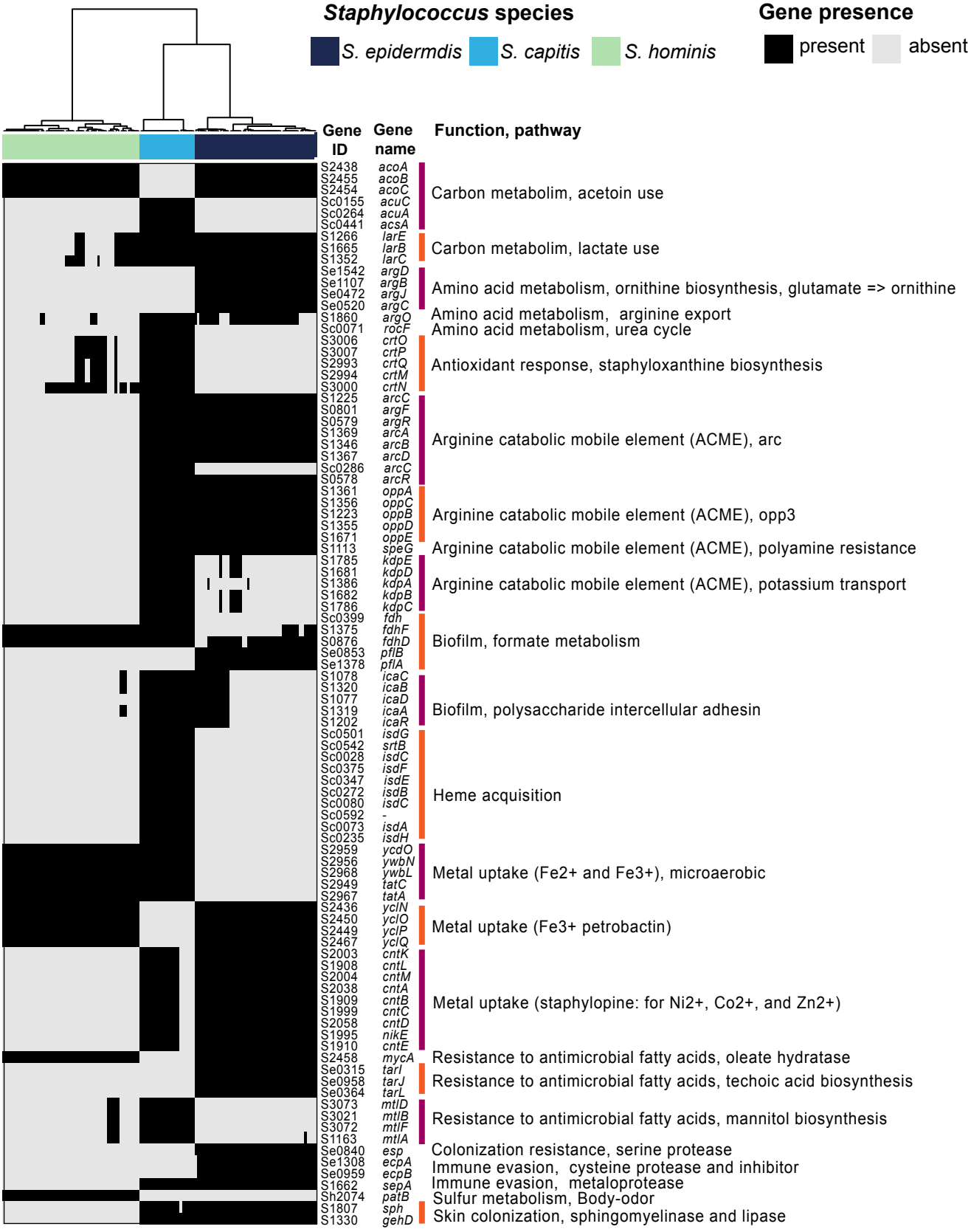

Figure S8. Species-restricted core genes relevant to skin colonization.

Presence/absence patterns of select genes/loci that encode a complete pathway (by themselves or along with genus-core genes) and have a known role in skin colonization are shown. Gene ID, gene names and function encoded by each gene is shown on the right. Hierarchical clustering (Euclidean distance; Ward) of species is shown as a dendrogram on top.

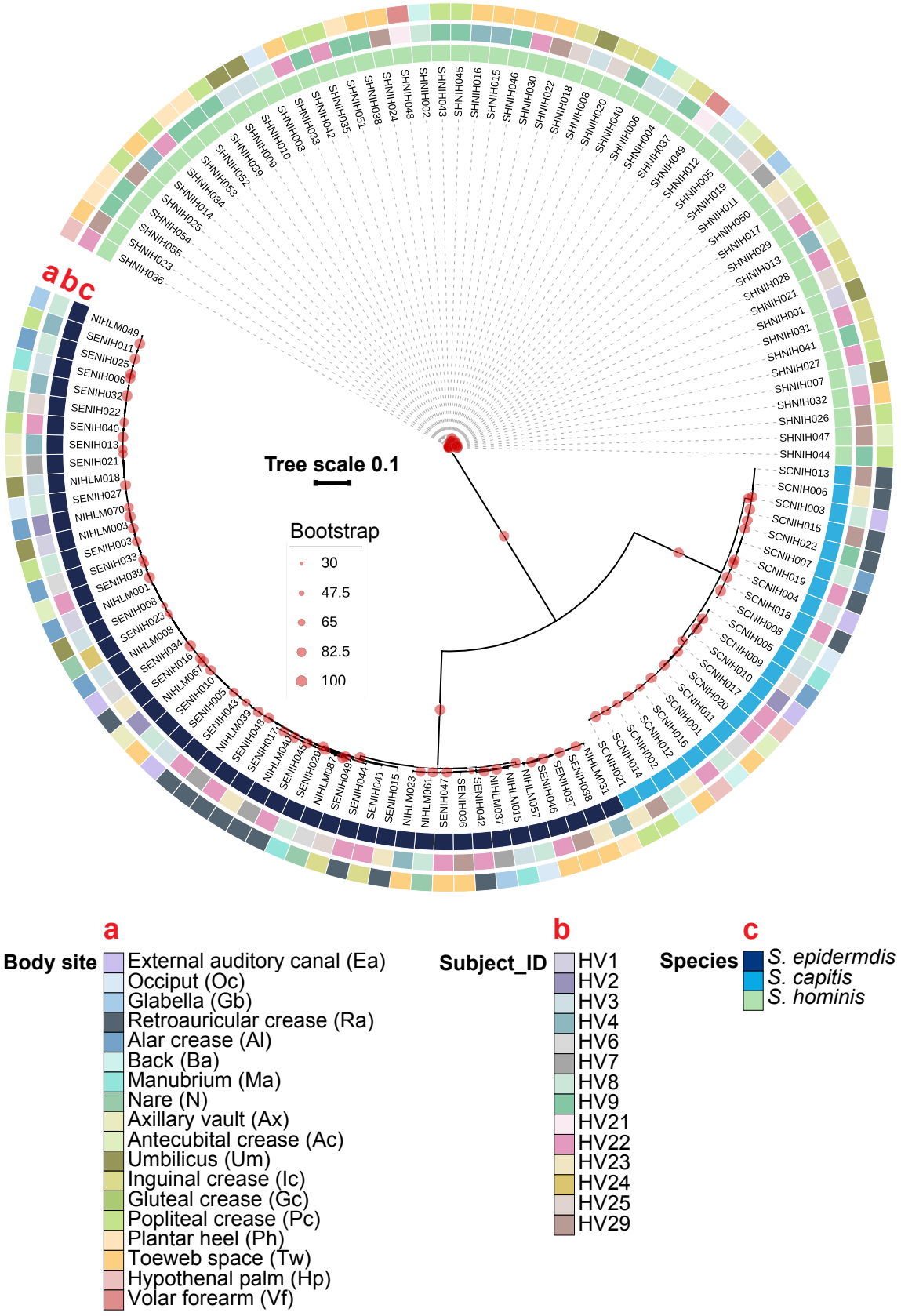

Figure S9. Phylogenetic tree of all 126 staphylococcal genomes based on the polymorphic sites detected in sequence alignment of select core genes (N = 665).

For each species-level phylo-genetic tree shown in Figure 3, the other two species served as an outgroup for rooting the tree based on this genus-level tree. Body site, healthy volunteer, and species of each genome are shown as sidebars. Refer to Fig. S1 for body site details.

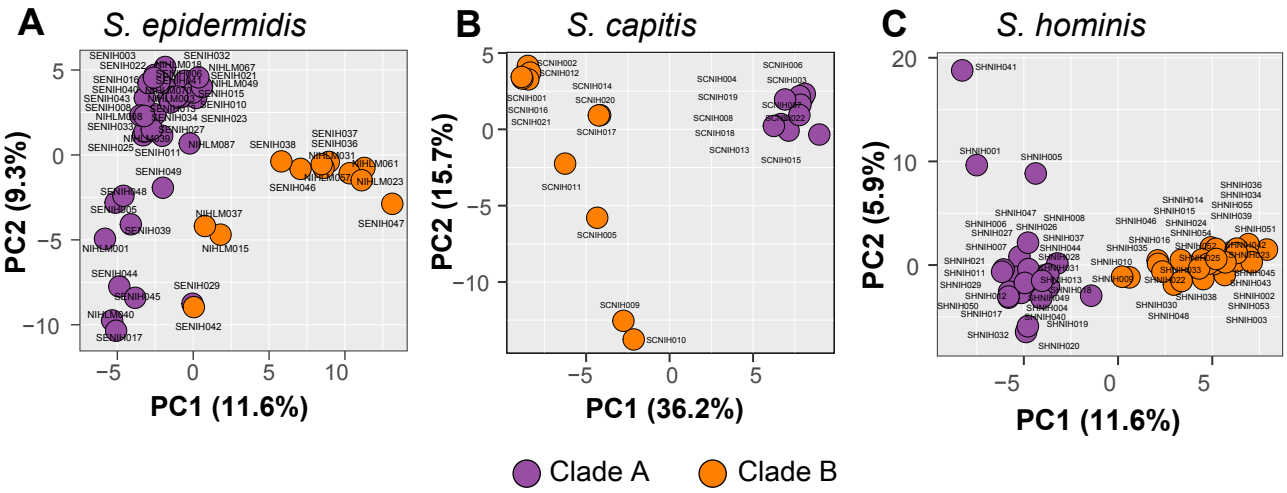

Figure S10. **Principal component analysis of conspecific genomes based on their gene presence/absence patterns.**

First two axes are plotted for each analysis, and account for 20.9%, 51.9% and 17.5% of the total variance for *S. epidermidis*, *S. capitis* and *S. hominis*, respectively. The first axis largely separates isolates by the two phylogenetic clades. Colors indicate the phylogenetic clades of isolates.

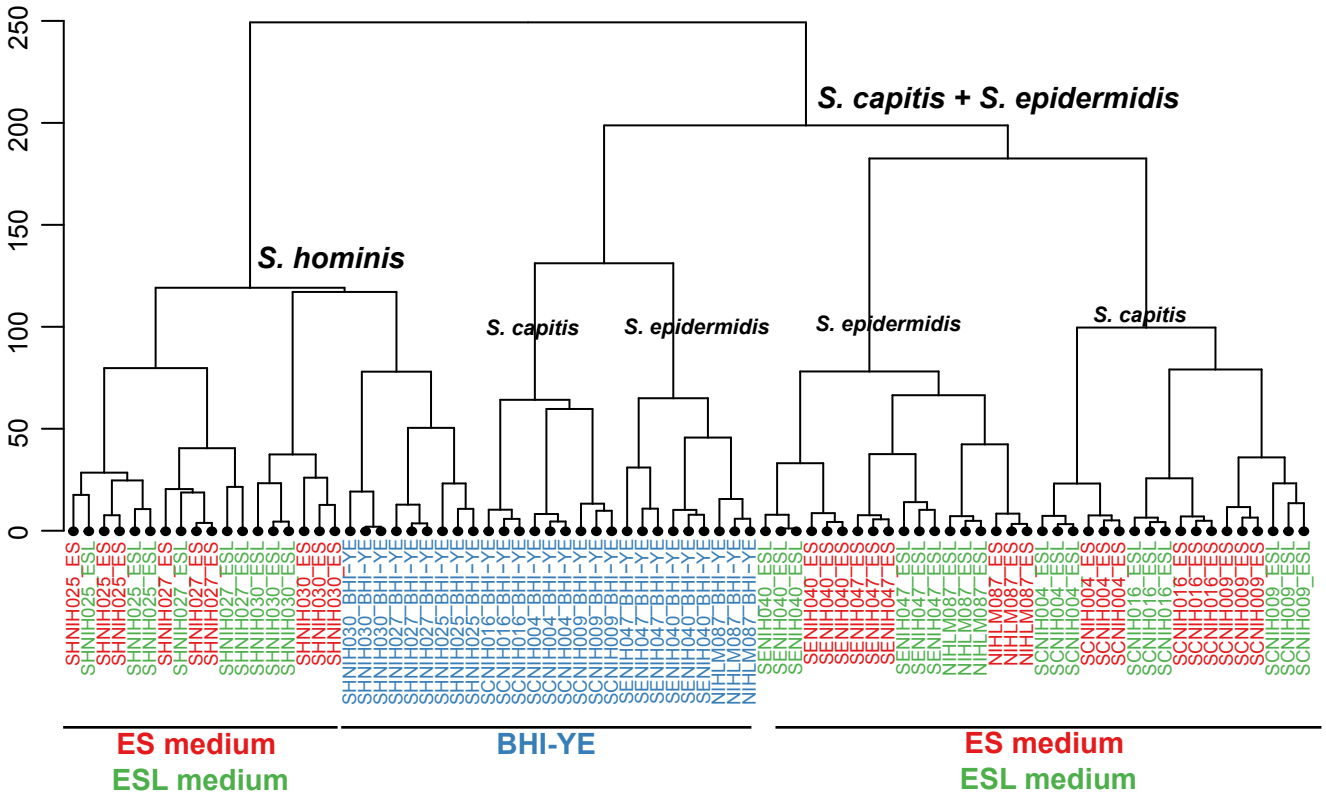

Figure S11 Hierarchal clustering of variance stabilizing transformation-normalized reads for 1647 genus-core genes from all RNA-seq samples (Euclidean distance; Ward).
The x-axis represents sample clusters, and the y-axis represents the distances between samples. Colors depict different growth media. Red: ES, Green: ESL, and Blue is BHI-YE. Samples clustered by species (shown as labels) and by growth medium.

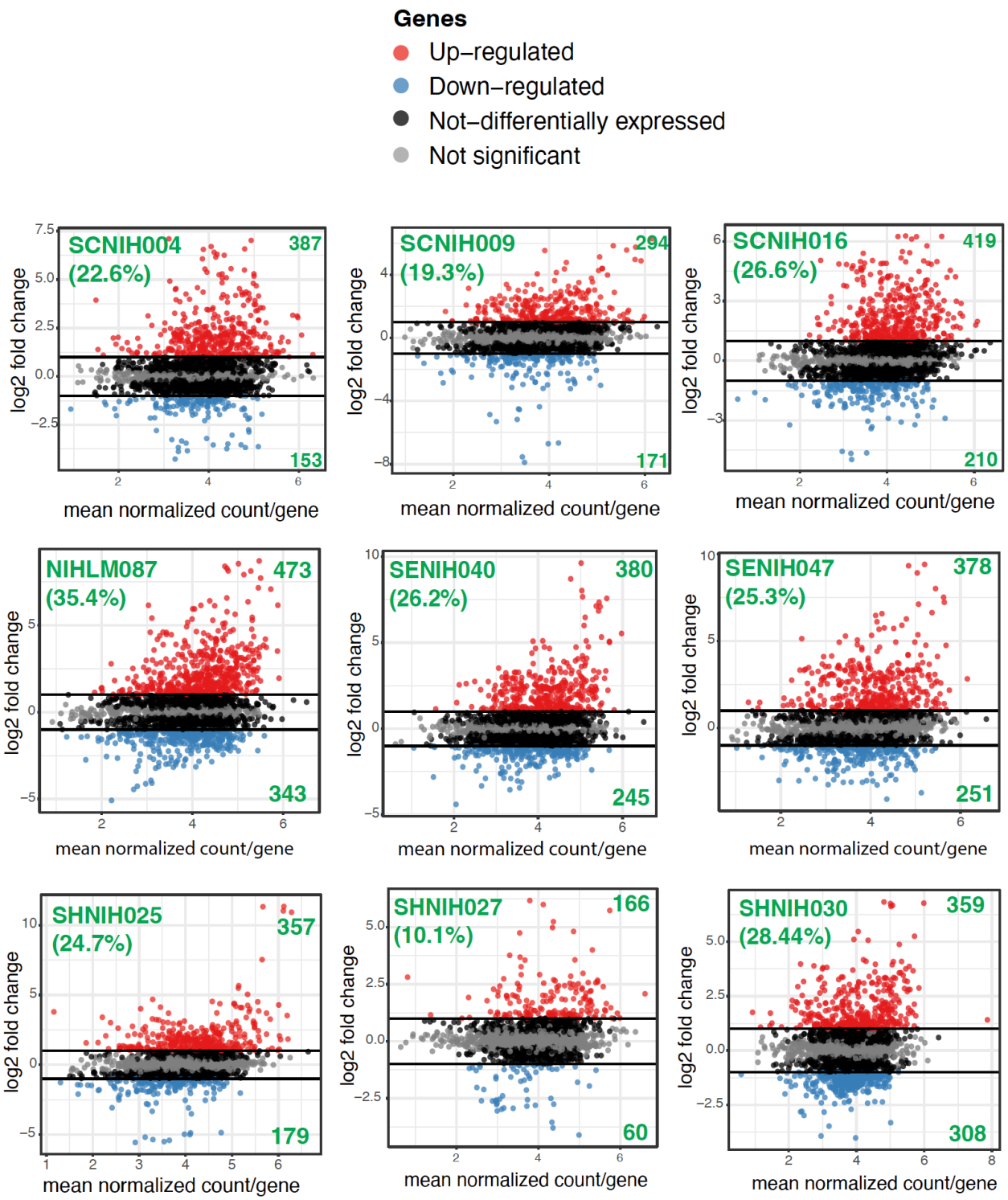

Figure S12 Differential gene expression plot for each isolate.
Each box represents the fold change in expression of each gene in the ES medium relative to BHI-YE for each isolate plotted against the mean normalized count per gene. Red dots represent upregulated genes, blue represent downregulated genes (≥ or ≤ 2-fold change, adjusted P value < 0.05, DESeq). The actual number of upregulated (top) and downregulated (bottom) genes is show in green. The numbers shown in green inside brackets placed below each isolate label represent the percentage of genes within a genome that were differentially regulated in the ES medium relative to BHI-YE.

Group Authors for NISC Comparative Sequencing Program:
Beatrice B Barnabas, Sean Black, Gerard G Bouffard, Shelise Y Brooks, Juyun Crawford, Holly Marfani, Lyudmila Dekhtyar, Joel Han, Shi-Ling Ho, Richelle Legaspi, Quino L Maduro, Catherine A Masiello, Jennifer C McDowell, Casandra Montemayor, James C Mullikin, Morgan Park, Nancy L Riebow, Karen Schandler, Brian Schmidt, Christina Sison, Sirintorn Stantripop, James W Thomas, Pamela J Thomas, Meghana Vemulapalli, Alice C Young