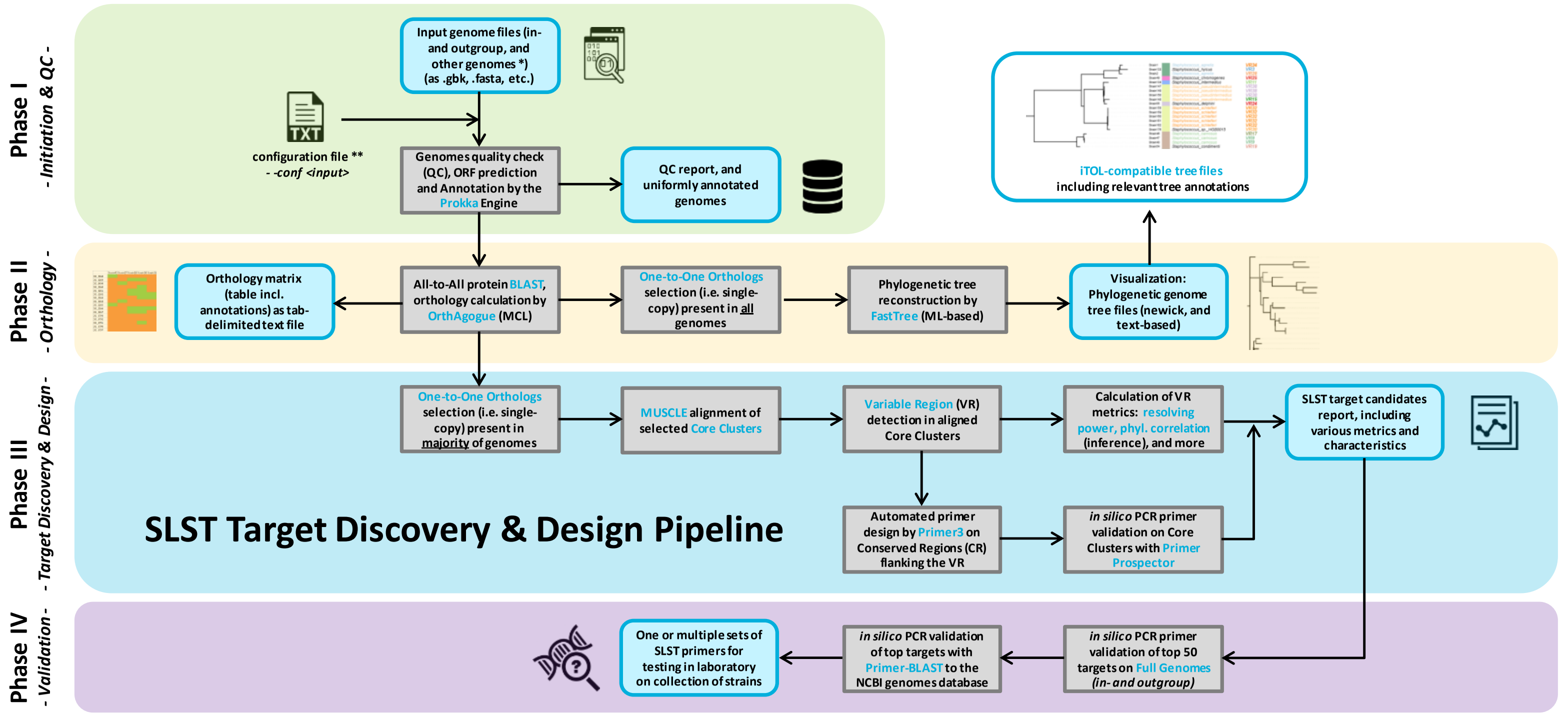
Supplementary Figures

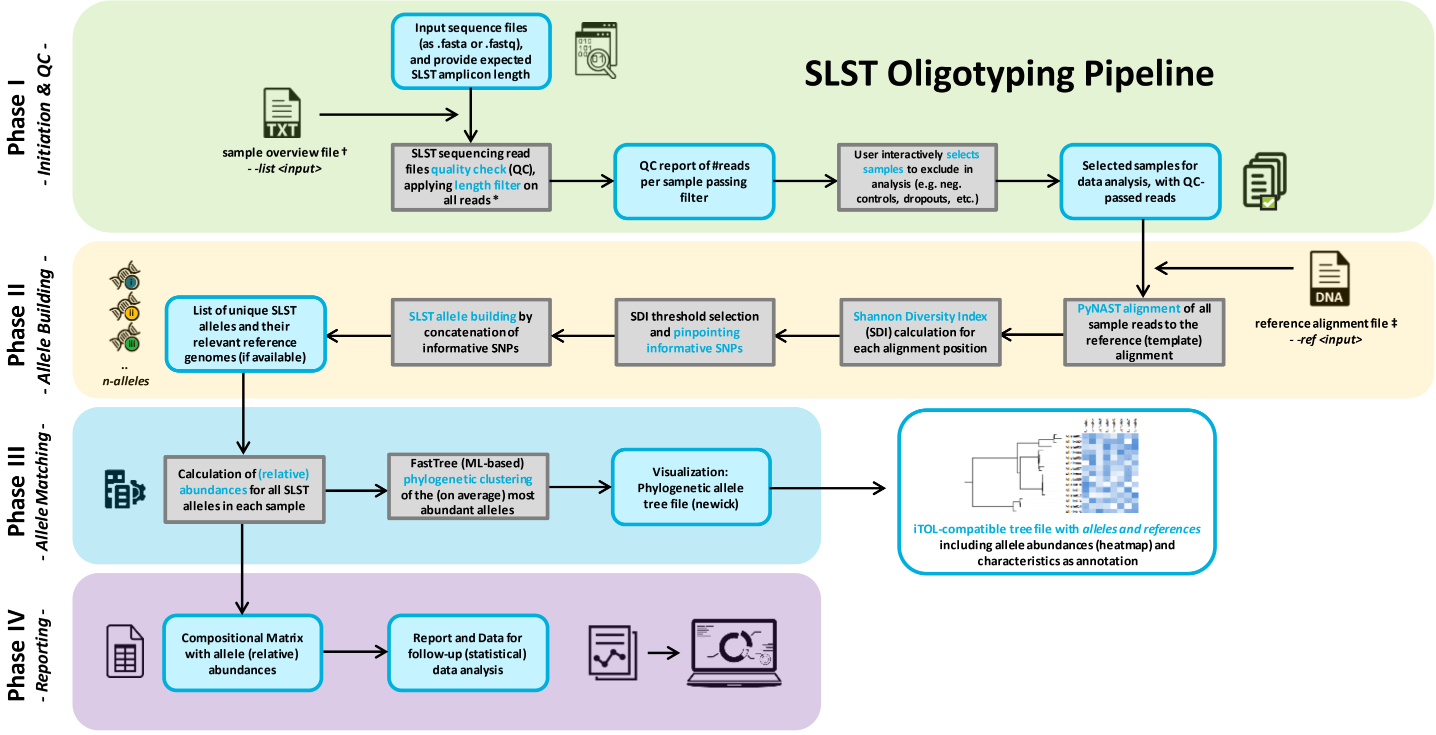


Phase I
- Initiation & QC -
Phase II
- Orthology -
Phase III
- Target Discovery & Design -
Phase IV
- Validation -
SLST Target Discovery & Design Pipeline
Input genome files (in- and outgroup, and other genomes *) (as .gbk, .fasta, etc.)
TXT
configuration file **
- -conf <input>
Genomes quality check (QC), ORF prediction and Annotation by the Prokka Engine
QC report, and uniformly annotated genomes
All-to-All protein BLAST, orthology calculation by OrthAgogue (MCL)
Orthology matrix (table incl. annotations) as tab-delimited text file
One-to-One Orthologs selection (i.e. single-copy) present in all genomes
Phylogenetic tree reconstruction by FastTree (ML-based)
Visualization: Phylogenetic genome tree files (newick, and text-based)
iTOL-compatible tree files including relevant tree annotations
One-to-One Orthologs selection (i.e. single-copy) present in majority of genomes
MUSCLE alignment of selected Core Clusters
Variable Region (VR) detection in aligned Core Clusters
Calculation of VR metrics: resolving power, phyl. correlation (inference), and more
SLST target candidates report, including various metrics and characteristics
Automated primer design by Primer3 on Conserved Regions (CR) flanking the VR
in silico PCR primer validation on Core Clusters with Primer Prospector
in silico PCR primer validation of top 50 targets on Full Genomes (in- and outgroup)
in silico PCR validation of top targets with Primer-BLAST to the NCBI genomes database
One or multiple sets of SLST primers for testing in laboratory on collection of strains

* *in- and outgroup genomes* will be fed into the pipeline, and will also be used for full-genome *in silico* PCR validation by Primer Prospector

** mandatory file containing user-provided information about input genomes: file name, annotation and status, i.e. how the pipeline should handle it (as either an 'in-/outgroup' genome, 'strain-of-interest' or other)

Figure S1. Overview of TaxPhlAn SLST Discovery and Design pipeline steps (Module A).


SLST Oligotyping Pipeline
Phase I
- Initiation & QC -
Phase II
- Allele Building -
Phase III
- Allele Matching -
Phase IV
- Reporting -
Input sequence files (as .fasta or .fastq), and provide expected SLST amplicon length
sample overview file †
- -list <input>
SLST sequencing read files quality check (QC), applying length filter on all reads *
QC report of #reads per sample passing filter
User interactively selects samples to exclude in analysis (e.g. neg. controls, dropouts, etc.)
Selected samples for data analysis, with QC-passed reads
reference alignment file ‡
- -ref <input>
PyNAST alignment of all sample reads to the reference (template) alignment
Shannon Diversity Index (SDI) calculation for each alignment position
SDI threshold selection and pinpointing informative SNPs
SLST allele building by concatenation of informative SNPs
List of unique SLST alleles and their relevant reference genomes (if available)
n-alleles
Calculation of (relative) abundances for all SLST alleles in each sample
FastTree (ML-based) phylogenetic clustering of the (on average) most abundant alleles
Visualization: Phylogenetic allele tree file (newick)
iTOL-compatible tree file with alleles and references including allele abundances (heatmap) and characteristics as annotation
Compositional Matrix with allele (relative) abundances
Report and Data for follow-up (statistical) data analysis

* only allowing reads within a three nucleotide (shorter) length of the expected SLST amplicon length (i.e. the reference SLST alignment length) to pass the filter

† file containing an overview of all sequencing samples, and whether or not to include them in the Shannon entropy analysis (all samples will be reported, except those deselected by the user, e.g. dropout samples)

‡ mandatory SLST alignment (template) file of all available reference genomes; is provided as output by the TaxPhlAn SLST Target Discovery & Design Pipeline

Figure S2. Overview of TaxPhlAn SLST Oligotyping pipeline steps (Module B).

Figures S3-S6. SLST candidates as predicted by TaxPhlAn correlates with the actual phylogenetic distances between genomes.

Shown in the PDF files (listed on the next page) is for each benchmark genome dataset a phylogenetic tree based on concatenated single nucleotide polymorphism (SNP) positions in the aligned core genome (one-to-one single-copy orthologous genes) of all genomes in that benchmark dataset (see Table 1 and Supplementary Tables S1-S4 for additional information on the benchmark datasets). One can appreciate that different variable regions (VR) of the SLST targets cluster almost entirely with the different bacterial clades based on full-genome information, and this was observed for all four bacterial datasets. Corresponding Spearman correlation values were on average 0.775 rho, and alternatively, Pearson's r correlation values were comparable (Supplementary Tables S6-S9). Other sequence distance methods such as DISTMAT [1] provided similar outcomes (Supplementary Tables S6-S9). Altogether, for automated SLST target prioritizing we decided to adopt ML-based Spearman correlations as we believe this is best practice, and data shows little difference from DISTMAT-based distances or correlation by Spearman. The phylogenetic trees are built with FastTree [2] which is maximum likelihood-based, and visualized with iTOL [3]. For the top SLST candidate of each benchmark genome dataset as analyzed with TaxPhlAn (Supplementary Tables S6-S9), the SLST variable regions (VR) are indicated with different colors and numbers (e.g. VR10), and the sequence alignment of the SLST regions is plotted on the right (aligned by MUSCLE, [4]). The number followed by an asterisk in between the brackets, such as (002*), represents the number of times that unique VR was found in the dataset. Blanc line means that based on the *in silico* PCR analysis by PrimerPropsector [5] no primer hit was available for that genome.

Files :
S3 Bifidobacterium-SLST-ID-630.24-34.21-21.iTOL.pdf (online) http://ederveen.science/Thesis/Chapter3/Fig-S3_Bifidobacterium-SLST-ID-630.24-34.21-21.iTOL.PRESUBMISSION.pdf
S4 Escherichia-Shigella-SLST-ID-1978.3-8.32-32.iTOL.pdf (online) http://ederveen.science/Thesis/Chapter3/Fig-S4_Escherichia-Shigella-SLST-ID-1978.3-8.32-32.iTOL.PRESUBMISSION.pdf
S5 Propionibacterium-SLST-ID-1414.3-7.9-9.iTOL.pdf (online) http://ederveen.science/Thesis/Chapter3/Fig-S5_Propionibacterium-SLST-ID-1414.3-7.9-9.iTOL.PRESUBMISSION.pdf
S6 Staphylococus-SLST-ID-1238.2-13.3-3.iTOL.pdf (online) http://ederveen.science/Thesis/Chapter3/Fig-S6_Staphylococus-SLST-ID-1238.2-13.3-3.iTOL.PRESUBMISSION.pdf

Figure S7. *Staphylococcus* full-genome phylogenetic tree projected with 16S clusters and SLST OG #1123 variable regions.

Shown in the PDF file is for the *Staphylococcus* dataset a phylogenetic tree based on concatenated single nucleotide polymorphism positions in the aligned core genome of all genomes in that benchmark dataset (n = 200). The phylogenetic tree is built with FastTree [2] which is maximum likelihood-based, and visualized with iTOL [3]. SLST candidate OG #1123 variable regions (VR) are indicated with different colors and numbers (e.g. VR10). In total, we identified 48 unique variable regions for OG #1123 (see Supplementary Table S10). If no VR was indicated (blanc line) this means that based on the *in silico* PCR analysis by PrimerPropsector [5] no primer hit was available for that genome (is only the case for Strain155). Colored blocks on the right of the phylogenetic tree represent unique 16S clusters based on full-length 16S sequences from these *Staphylococcus* genomes clustered on 99.7% identity (see '*Selection of representative Staphylococcus genomes for TaxPhlAn target discovery*' in main Methods for more details).

File :
S7 SLST-Staphylococcus-OG#1123.pdf (online) http://ederveen.science/Thesis/Chapter3/Fig-S7_SLST-Staphylococcus-OG1123.PRESUBMISSION.pdf

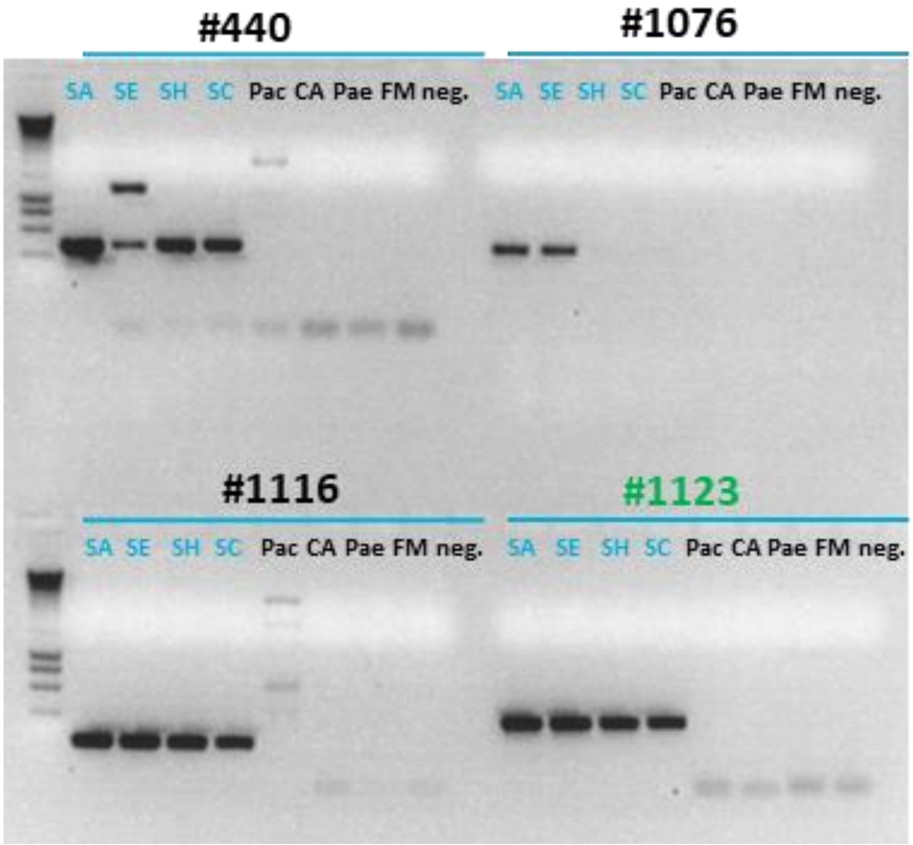

#440
SA SE SH SC Pac CA Pae FM neg.
#1076
SA SE SH SC Pac CA Pae FM neg.
#1116
SA SE SH SC Pac CA Pae FM neg.
#1123
SA SE SH SC Pac CA Pae FM neg.

Figure S8. PCR validation of SLST candidates primer sets with *Staphylococcus* species and common skin commensals.

SLST candidates (see Supplementary Table S13) OG #1116 and #440 showed cross-reactivity with *P. acnes*, in addition #440 had an off-target for *S. epidermidis*. OG # 1076 did not recognize *S. hominis* and S*. capitis*. In conclusion, OG #1123 was the perfect candidate here with recognition of all Staphylococci and without cross-reactivity with common skin commensals. DNA ladder used: 200bp. Note that the lowest slightly visible bands are primer-dimers. SA: *S. aureus*, SE: *S. epidermidis*, SH: *S. hominis*, SC: *S. capitis*, Pac: *Propionibacterium acnes*, CA: *Corynebacterium aurimucosum*, Pae: *Pseudomonas aeruginosa*, FM: *Finegoldia magna*, neg: negative control with sterile water. The SLST OG #1123 (dark green) is our final SLST target which survived all *in silico* and lab validations, and with which our study was performed. OG #1123 is predicted to encode for 30S ribosomal protein S11.

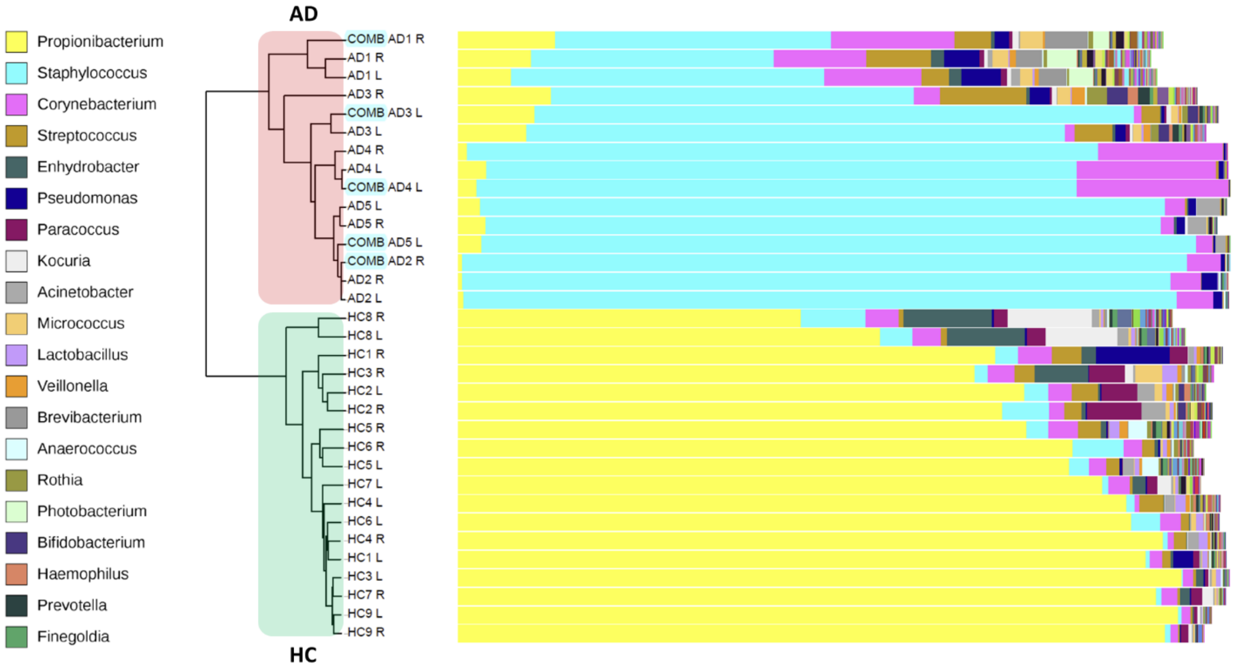


Figure S9. Skin microbiota composition in healthy controls and AD patients: *Staphylococcus* dominates skin of AD patients at the expense of *Propionibacterium*.

Each leaf of the tree represents a single sample. Samples were clustered based on beta diversity ("between-sample distance") using weighted UniFrac as a distance measure and hierarchical UPGMA as a clustering method. Vertical bars show the relative abundance microbiota composition on the genus level (reads that could not be classified up to this level are in white), although for clustering the full microbiome information was used. The 20 most dominant genera are shown in the legend, and are sorted from high to low. Colored sample labels represent sample classes: atopic dermatitis (AD) patients in red and healthy control (HC) volunteers in green. In addition, samples marked in blue (COMB) are technical replicated for which 16S amplicons were mixed with SLST amplicons of the same sample. The figure was generated with the interactive tree of life (iTOL) program [3].

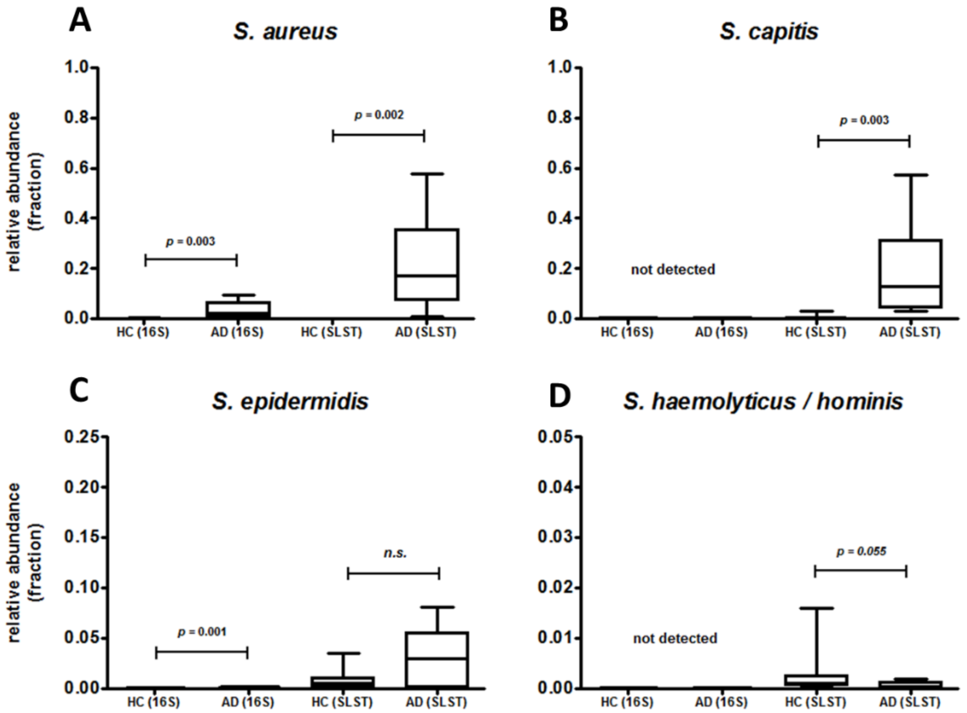


Figure S10. *Staphylococcus* species are increased in AD in comparison to HC.

Plotted in bar-graphs are the relative abundance levels of *Staphylococcus* species (**A**) *S. aureus*, (**B**) *S. capitis*, (**C**) *S. epidermidis* and (**D**) *S. haemolyticus / hominis*, as measured by 16S or SLST. AD = atopic dermatitis. HC = healthy control. We mainly identified taxa of *S. aureus* and *S. capitis*, which were significantly increased in AD, in comparison to taxa of *S. epidermidis* and a cluster of *S. haemolyticus / hominis*, which were associated with healthy skin. The species-level SLST relative abundances plotted here have, for each sample, been corrected for the corresponding relative abundance of *Staphylococcus* (genus-level) for that sample based on 16S. Boxplots as median with interquartile range and whiskers from min to max.

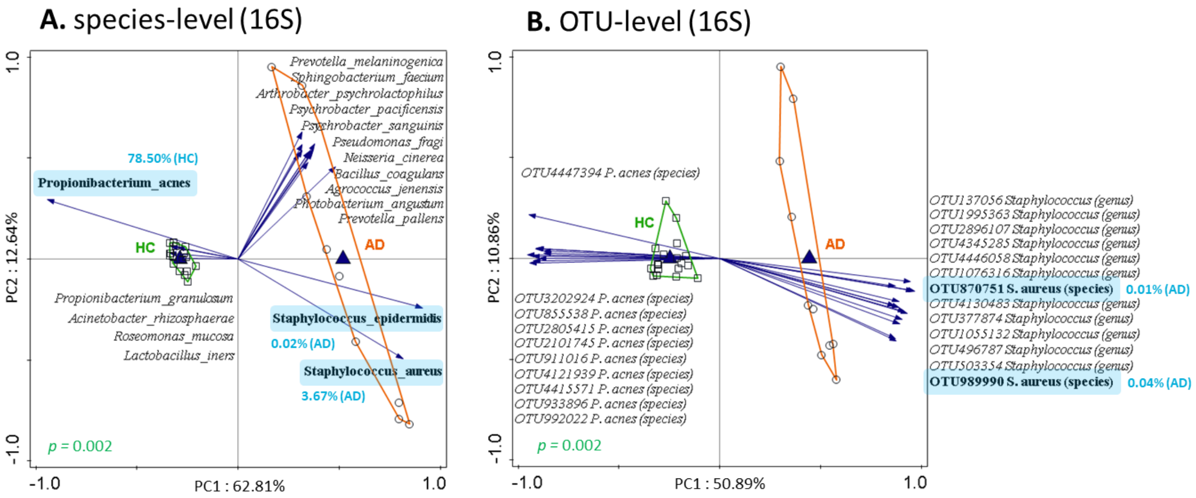


Figure S11. Strong under-assignment of 16S *Staphylococcus* taxa below the level of genus.

The figures show a redundancy analysis (RDA) biplot of 16S data on (**A**) species-level and (**B**) OTU-level. Triangles are the centroids of the study sample groups: HC (green) and AD (orange). The blue arrows are the best-fitting bacterial taxa (names in italic or bold) with a minimum of 7.5% (A) or 75% (B) horizontal fit value, i.e. taxa that best explain the differences between the sample groups. The horizontal axis maximizes the variation in sample groups (in contrast to a principal component analysis plot, where the variation between individual samples is maximized). In RDA, samples (also) separated in the vertical direction indicates that this separation is (also) driven by other factors than the primary contrast, such as by individuality. The difference in microbiota is significant on both taxonomical level (according to a permutation test, which was corrected for individual; *p* = 0.002). Blue percentage numbers represent the on average relative abundance of that response variable for HC or AD; this is to show the relatively low relevance of *S. aureus*- and *S. epidermidis*-assigned genera or OTU in separation between HC and AD based on 16S data alone. SA = *S. aureus*-like taxon (allele). SC = *S. capitis*-like taxon (allele).

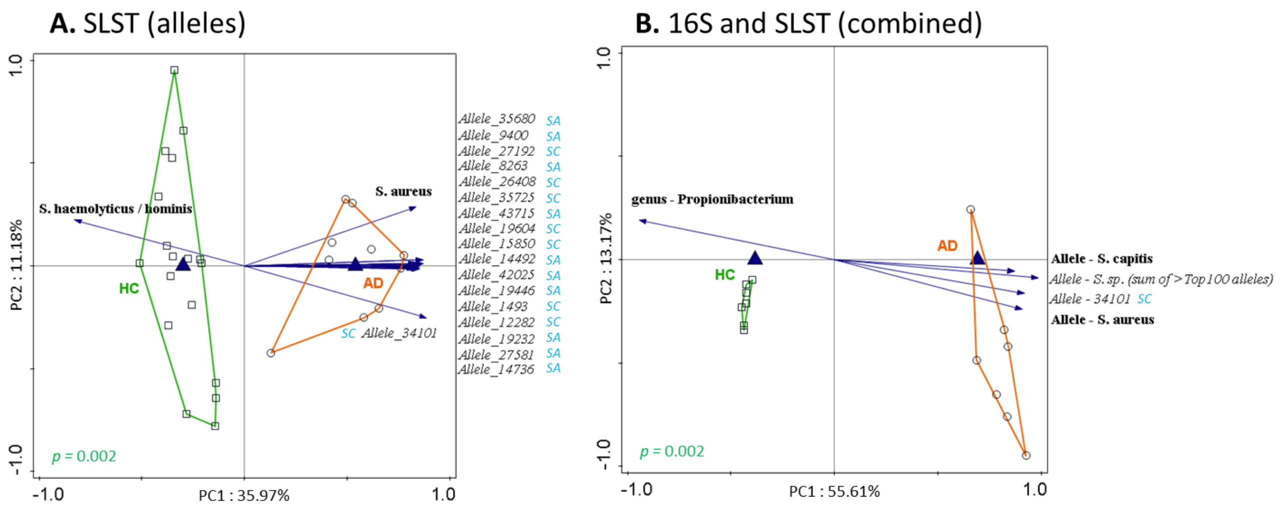


Figure S12. The below genus-level bacterial entities discriminating HC and AD individuals is mainly explained by *S. aureus and S. capitis* species.

The figures show a redundancy analysis (RDA) biplot of (**A**) *Staphylococcus* SLST allele data, and of (**B**) 16S data combined with *Staphylococcus* SLST allele information. In other words, SLST data in (B) has been appended (and corrected) with available 16S data. Triangles are the centroids of the study sample groups: HC (green) and AD (orange). The blue arrows are the best-fitting bacterial taxa (names in italic or bold) with a minimum of 68% horizontal fit value, i.e. taxa that best explain the differences between the sample groups. The horizontal axis maximizes the variation in sample groups (in contrast to a principal component analysis plot, where the variation between individual samples is maximized). In RDA, samples (also) separated in the vertical direction indicates that this separation is (also) driven by other factors than the primary contrast, such as by individuality. The difference in microbiota is significant with both 16S uncorrected (A) and 16S corrected (B) data (according to a permutation test, which was corrected for individual; *p* = 0.002). SA = *S. aureus*-like taxon (allele). SC = *S. capitis*-like taxon (allele).

References Supplementary Figures

1. Rice, P., I. Longden, and A. Bleasby, *EMBOSS: the European Molecular Biology Open Software Suite.* Trends Genet, 2000. **16**(6): p. 276-7.
2. Price, M.N., P.S. Dehal, and A.P. Arkin, *FastTree: computing large minimum evolution trees with profiles instead of a distance matrix.* Mol Biol Evol, 2009. **26**(7): p. 1641-50.
3. Letunic, I. and P. Bork, *Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees.* Nucleic Acids Res, 2016. **44**(W1): p. W242-5.
4. Edgar, R.C., *MUSCLE: multiple sequence alignment with high accuracy and high throughput.* Nucleic Acids Res, 2004. **32**(5): p. 1792-7.
5. Walters, W.A., et al., *PrimerProspector: de novo design and taxonomic analysis of barcoded polymerase chain reaction primers.* Bioinformatics, 2011. **27**(8): p. 1159-61.