nature biotechnology

Supplementary information

https://doi.org/10.1038/s41587-025-02797-4

Skin metatranscriptomics reveals a landscape of variation in microbial activity and gene expression across the human body

In the format provided by the authors and unedited

Supplementary Note 1: Analysis of viral reads from skin metatranscriptomes

Only a very small percentage (median 0.01-0.07%) of all microbial reads (bacterial, archaeal, fungal and viral) were classified as viral in our metagenomic and metatranscriptomic datasets across all skin sites (**Supplementary Note Figure 1**). Based on our default thresholds, these taxonomic assignments were filtered out by our standard workflow (**Methods**). Using filters recommended by Breitwieser et al[1] (≥10 read pairs and coverage≥1,000 distinct Kraken2 minimizers per taxa), some of the viral taxa were recovered. The most abundant viruses identified in our skin metatranscriptomes belonged to the *Pahexavirus* genus, which includes diverse species of *Propionibacterium* phages (**Supplementary Note Figure 2**). *Gamma-* and *beta-papillomaviruses* were also sporadically observed, consisting with their detection on the skin of some asymptomatic individuals[2].

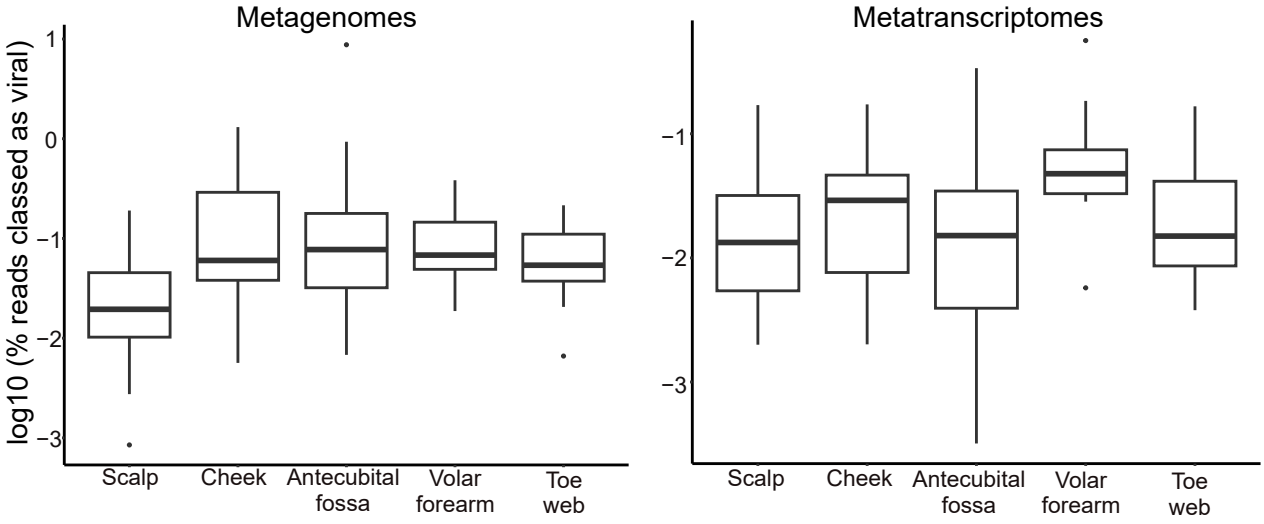


Figure 1: Boxplots showing percentage of microbial reads classified as viruses across body sites for the full cohort of metagenomes and metatranscriptomes (n=102 samples)

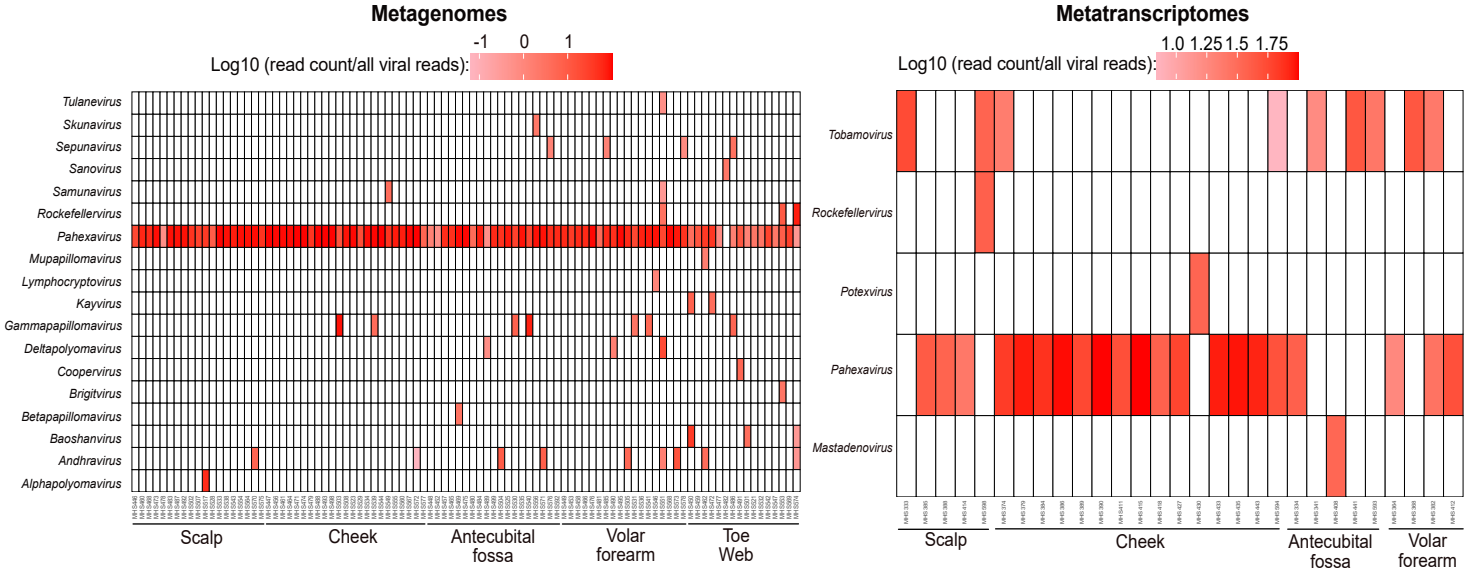


Figure 2: Heatmaps of relative abundances for various viral genera, normalized to the total number of viral reads in each sample. Each viral genus shown here had ≥10 read

pairs and ≥1,000 distinct Kraken2 minimizers within the sample. Only samples with at least one positive viral genus are shown here.

Supplementary Note 2: Decontamination of human reads

While mapping to a single reference (hg38) can be insufficient for removing human reads, a two-stage filtering approach using mapping followed by classification with a Kraken2 database containing the human genome can mitigate this problem[3]. This is the approach that we adopted for this manuscript.

To estimate the number of misclassified reads in our dataset after this two-stage approach, we extracted all reads classified as microbial by Kraken2 and mapped them to the human pangenome (hprc-v1.1-mc-chm13) from the Human Pangenome Reference Consortium[4] using VG Giraffe[5]. Across samples in the full cohort (n=102 metagenomes), only a very low percentage of Kraken2 classified microbial reads were mapped to the human pangenome (median 0.008%, range 0.0006%-0.17%). We further confirmed that the impact on the abundances of each species in each sample was correspondingly also minimal (<1% of reads; **Supplementary Note Figure 3**). Future skin metagenomic and metatranscriptomic workflows could benefit from more comprehensive host read removal using the human pangenome but splice-aware mapping of short RNA-seq reads to the human pan-transcriptome[6] is a relatively recent development and will require integration into state-of-the-art microbiome analysis workflows.

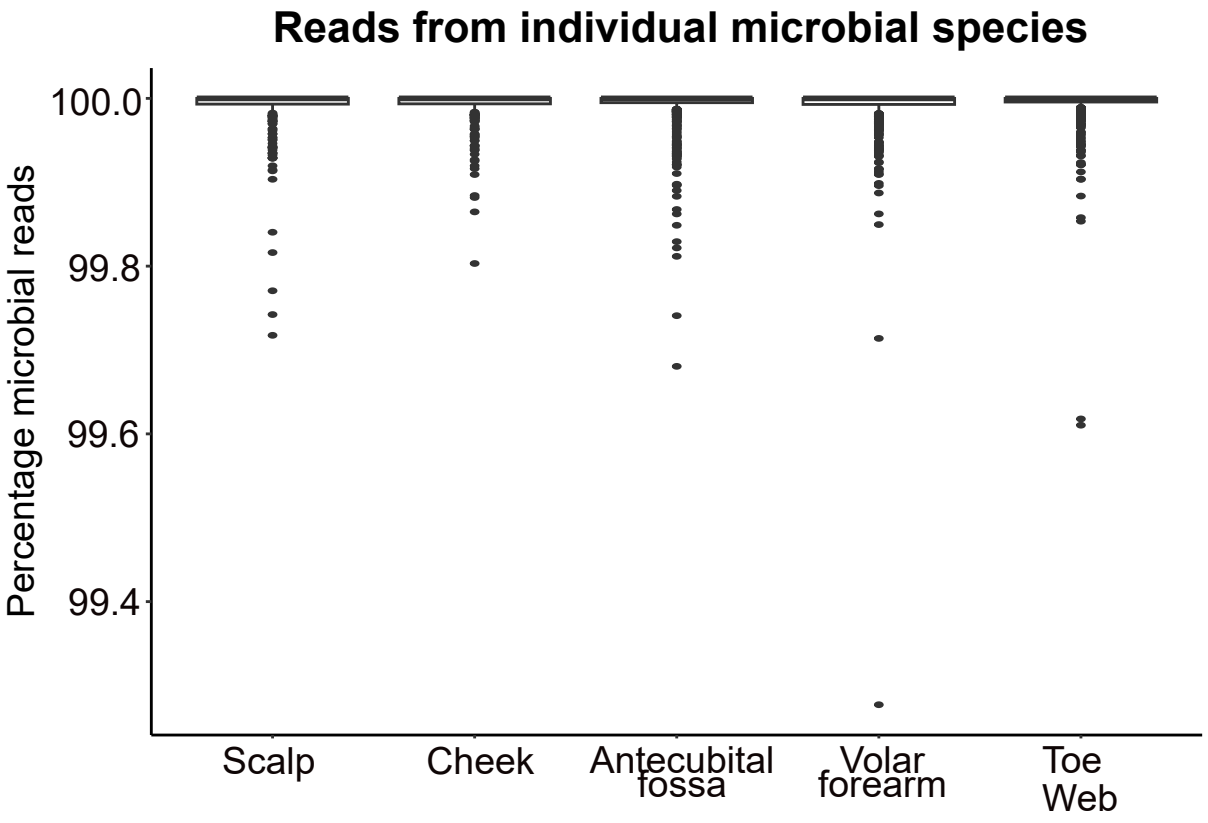


Figure 3: Boxplot showing percentage of reads for individual species which were not mapped to the human pangenome, across the full cohort (n =102 metagenomes). Only species in a sample with at least 100 reads were considered for this analysis.

Supplementary Note 3: Concordance of results from metatranscriptomic and metagenomic analysis workflows

It is generally important to minimize differences between the bioinformatic workflows for comparing metagenomic versus metatranscriptomic abundances. However, some aspects of RNA read processing such as computational rRNA filtering and spliced read alignment are not applicable for metagenomes. Nonetheless, to examine if workflow differences were primarily responsible for the differences observed in taxonomic composition between metagenomes and metatranscriptomes, we subjected metagenomes from the full cohort (n=102 libraries) to the exact same bioinformatic workflow as the metatranscriptomes. We found very high concordance between species-level metagenomic abundances between the two different workflows (R=1; **Supplementary Note Figure 4**) and correspondingly the striking divergences between metatranscriptomes and metagenomes remained.

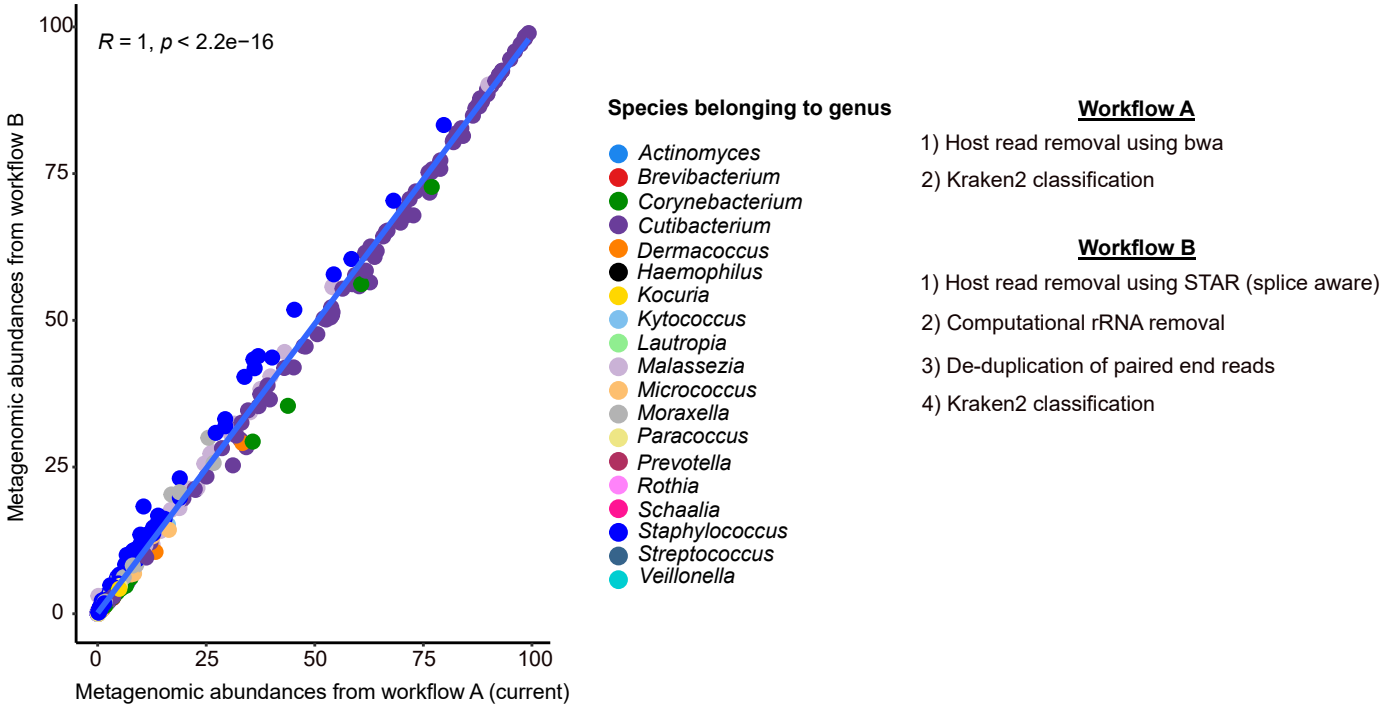


Figure 4: Scatter plot showing correlation of species-level metagenomic abundances from the full cohort (n=102 metagenomes) between two different bioinformatic workflows: workflow A (currently used) and workflow B.

References

1. Breitwieser, F.P., Baker, D.N., and Salzberg, S.L. (2018). KrakenUniq: Confident and fast metagenomics classification using unique k-mer counts. Genome Biol *19*, 1–10. https://doi.org/10.1186/S13059-018-1568-0.

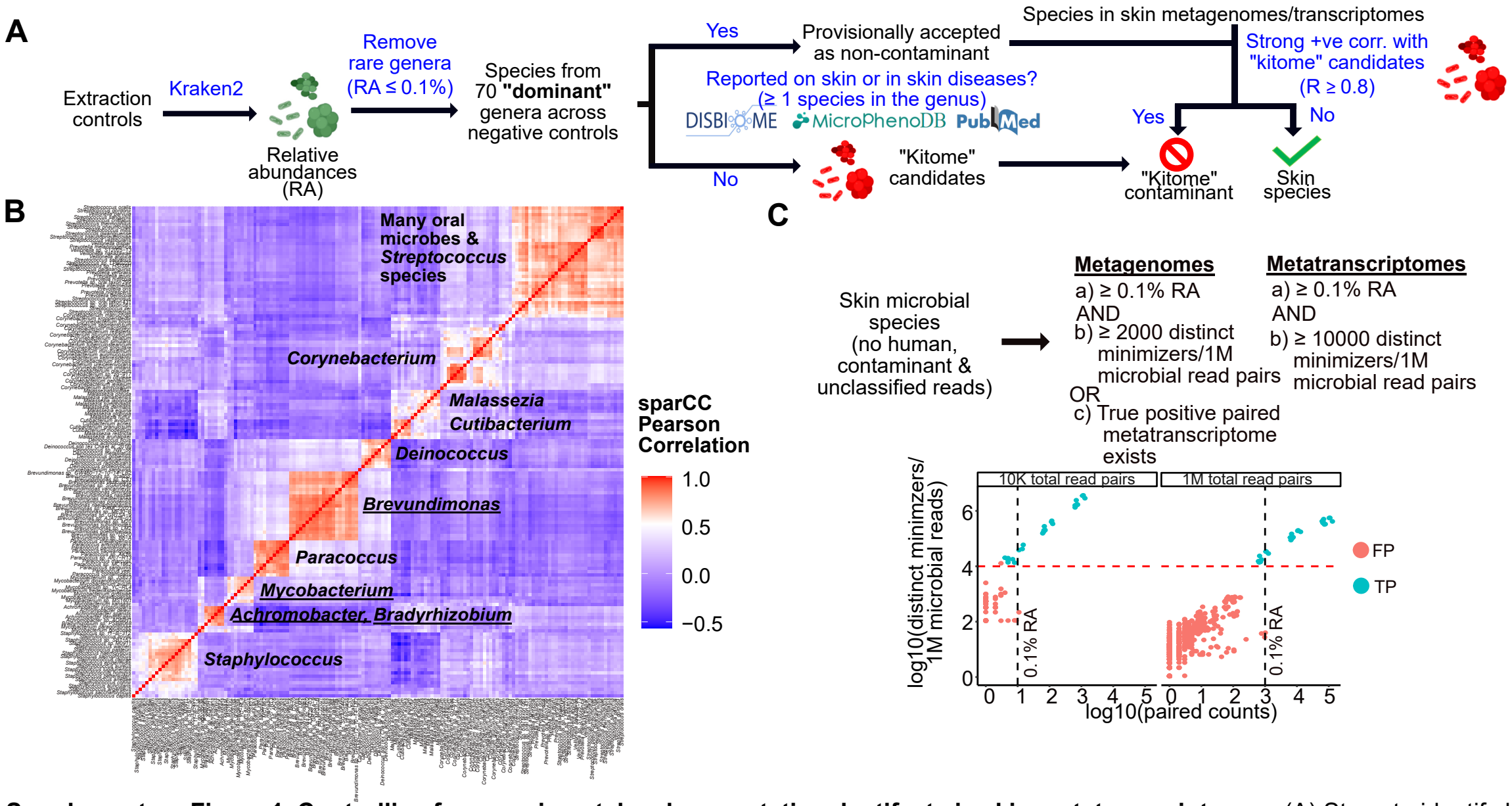
2. Foulongne, V., Sauvage, V., Hebert, C., Dereure, O., Cheval, J., Gouilh, M.A., Pariente, K., Segondy, M., Burguière, A., Manuguerra, J.C., et al. (2012). Human Skin Microbiota: High Diversity of DNA Viruses Identified on the Human Skin by High Throughput Sequencing. PLoS One *7*, e38499. https://doi.org/10.1371/JOURNAL.PONE.0038499.

3. Gihawi, A., Ge, Y., Lu, J., Puiu, D., Xu, A., Cooper, C.S., Brewer, D.S., Pertea, M., and Salzberg, S.L. (2023). Major data analysis errors invalidate cancer microbiome findings. mBio *14*. https://doi.org/10.1128/MBIO.01607-23.

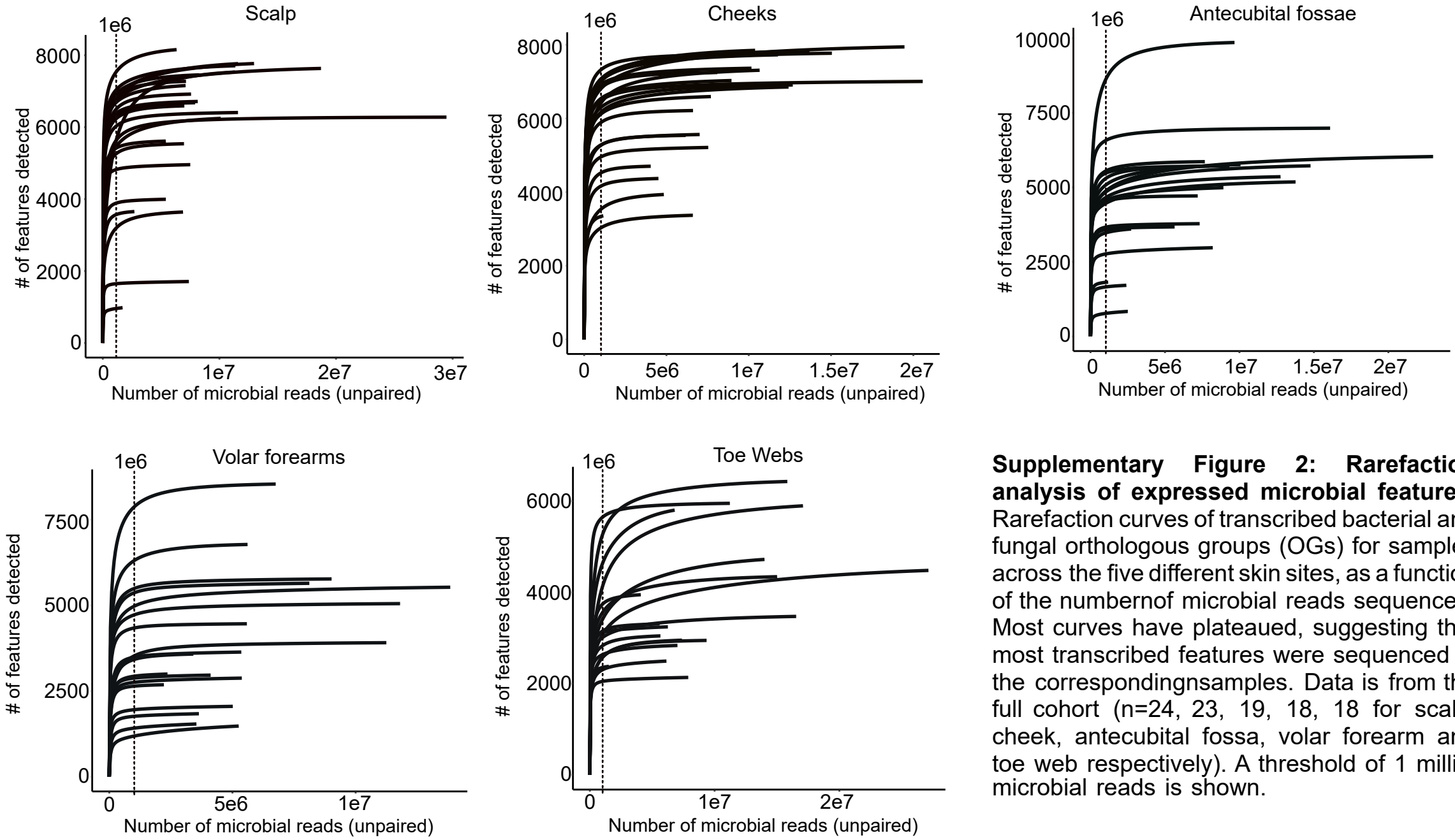
4. Liao, W.W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G., Lu, S., Lucas, J.K., Monlong, J., Abel, H.J., et al. (2023). A draft human pangenome reference. Nature 2023 617:7960 *617*, 312–324. https://doi.org/10.1038/s41586-023-05896-x.

5. Sirén, J., Monlong, J., Chang, X., Novak, A.M., Eizenga, J.M., Markello, C., Sibbesen, J.A., Hickey, G., Chang, P.C., Carroll, A., et al. (2021). Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. Science (1979) *374*. https://doi.org/10.1126/SCIENCE.ABG8871.

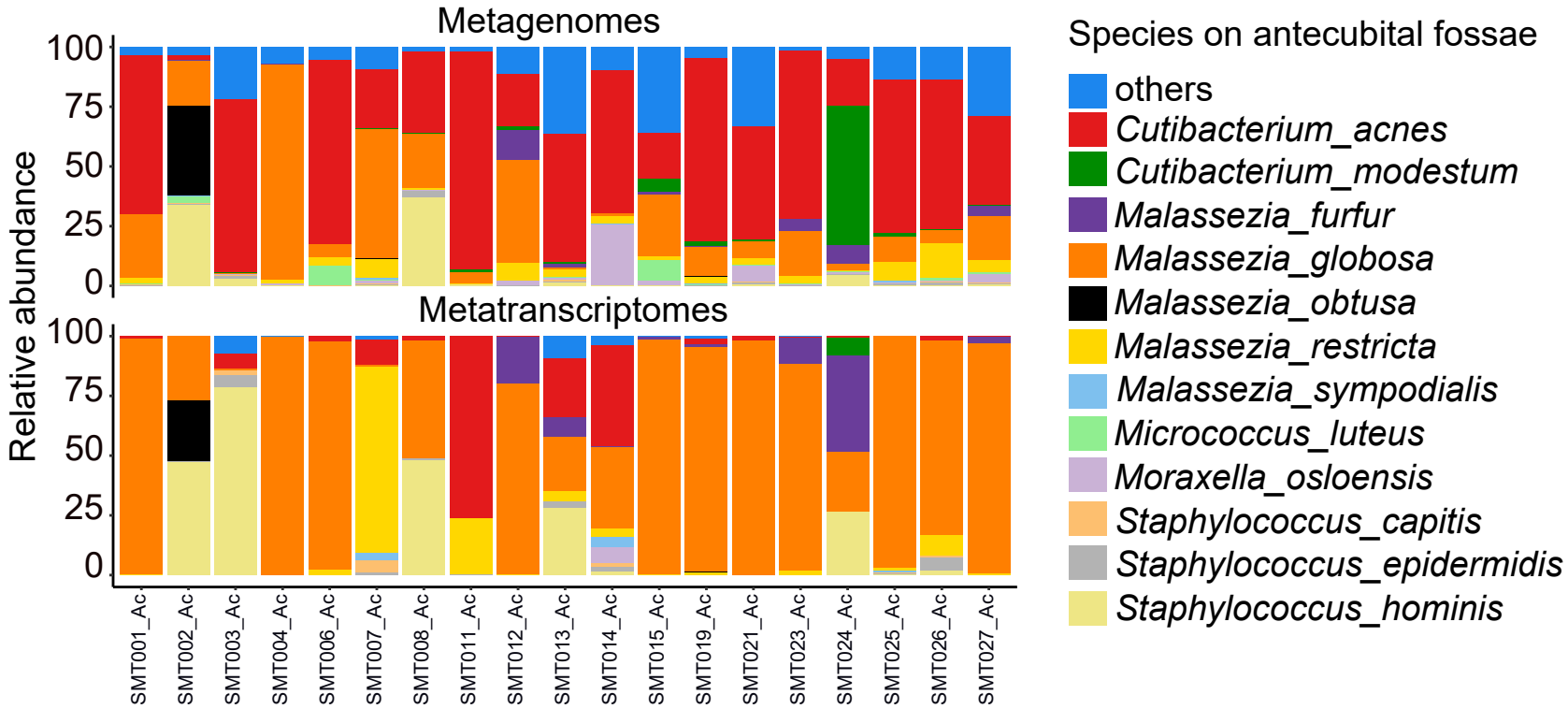
6. Sibbesen, J.A., Eizenga, J.M., Novak, A.M., Sirén, J., Chang, X., Garrison, E., and Paten, B. (2023). Haplotype-aware pantranscriptome analyses using spliced pangenome graphs. Nat Methods *20*, 239–247. https://doi.org/10.1038/S41592-022-01731-9



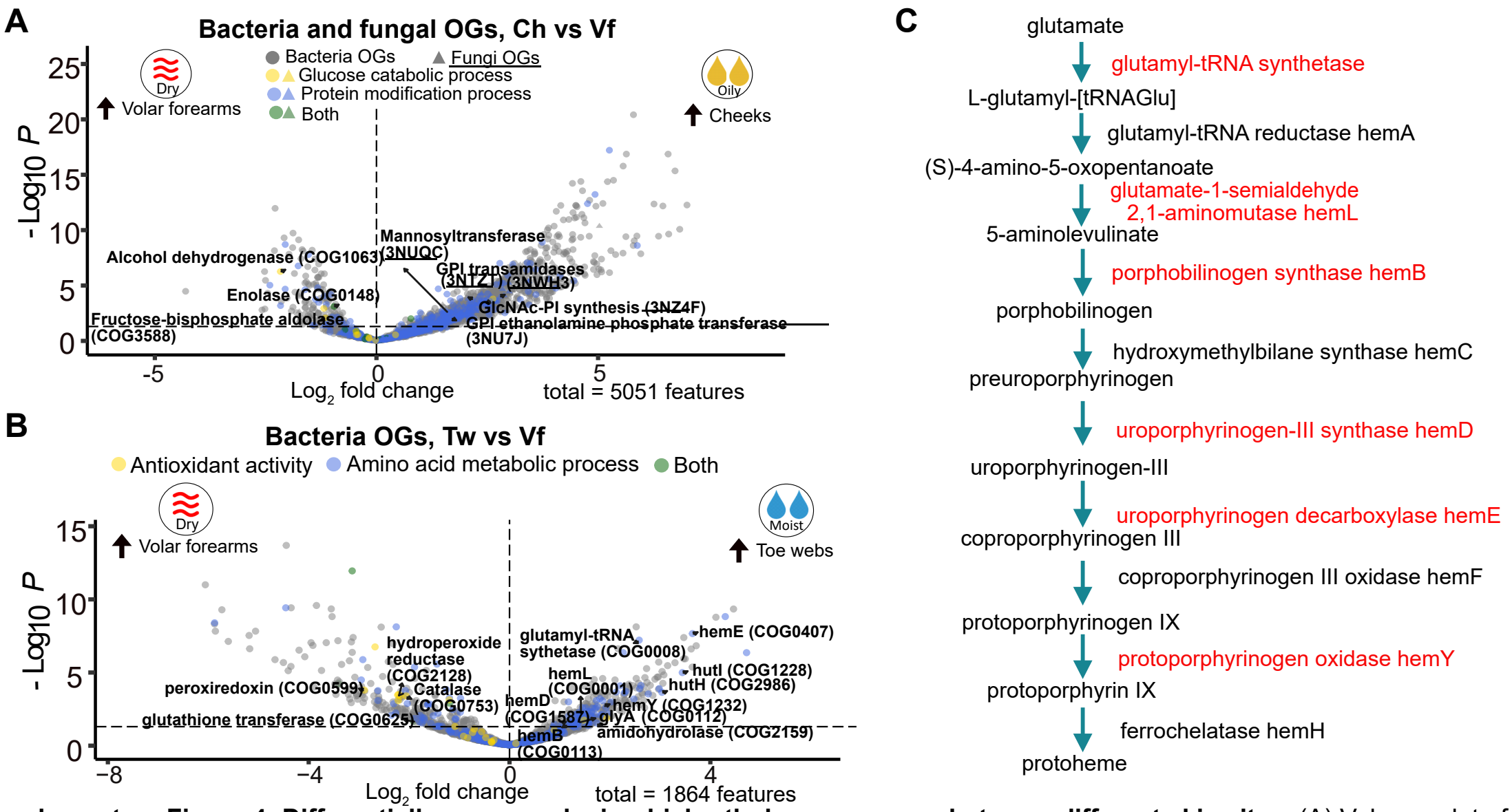
Supplementary Figure 1: Controlling for experimental and computational artifacts in skin metatranscriptomes. (A) Steps to identify kit contaminants. Created in BioRender. Ng, A. (2025) https://BioRender.com/ejgve5y. (B) Heatmap showing metagenomic abundances of contaminant genera (underlined) are not correlated with skin and oral microbes, facilitating their removal. (C) Criteria for differentiating true positive (TP) from false positive (FP) species assignments using thresholds for relative abundance (black dotted line) and for log10 distinct fminimizers for a given species per 1M microbial reads (red dotted line), for different library sizes (10K and 1M total read pairs).



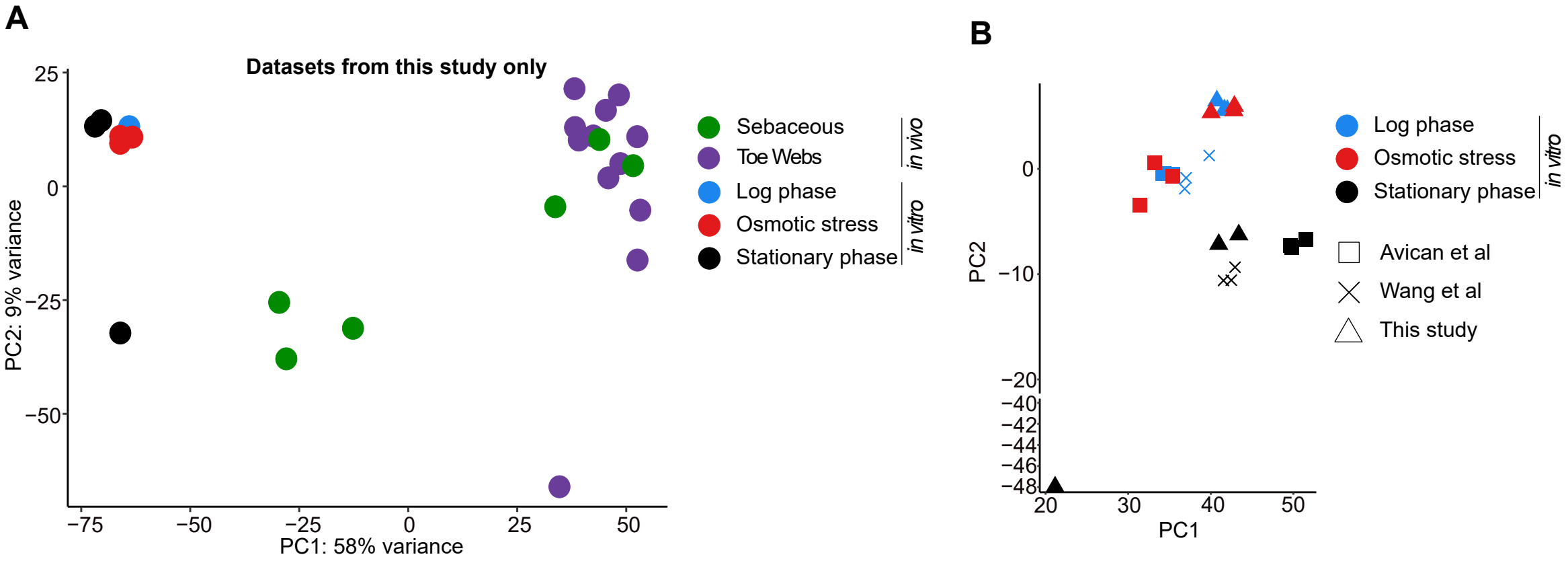
Supplementary Figure 2: Rarefaction analysis of expressed microbial features. Rarefaction curves of transcribed bacterial and fungal orthologous groups (OGs) for samples across the five different skin sites, as a function of the numbernof microbial reads sequenced. Most curves have plateaued, suggesting that most transcribed features were sequenced in the correspondingnsamples. Data is from the full cohort (n=24, 23, 19, 18, 18 for scalp, cheek, antecubital fossa, volar forearm and toe web respectively). A threshold of 1 million microbial reads is shown.



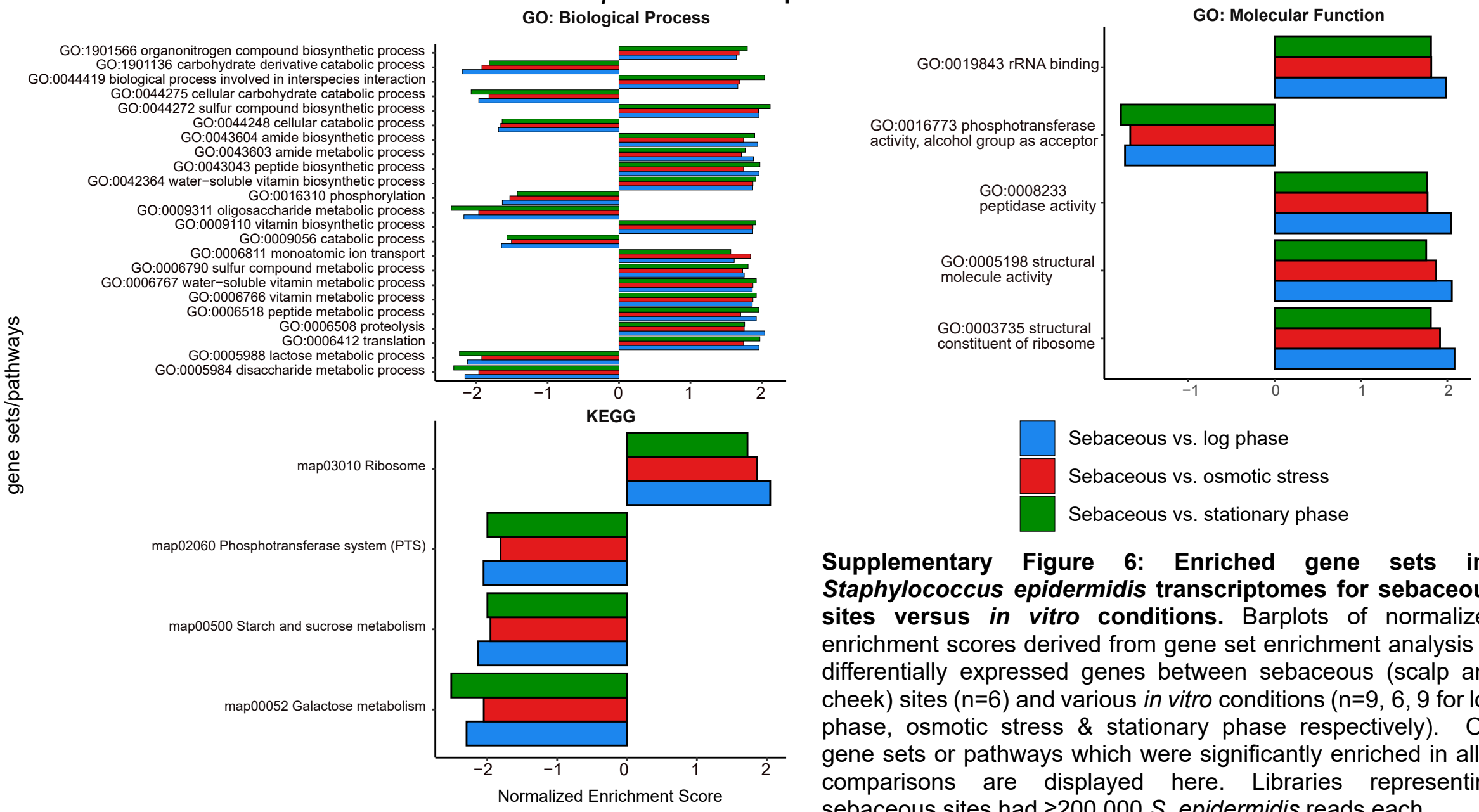
Supplementary Figure 3: Comparison of metagenomic and metatranscriptomic species-level relative abundances in the antecubital fossae across individuals. Stacked bar plots showing the most abundant species for matched skin metagenomes and metatranscriptomes across individuals (n=19).



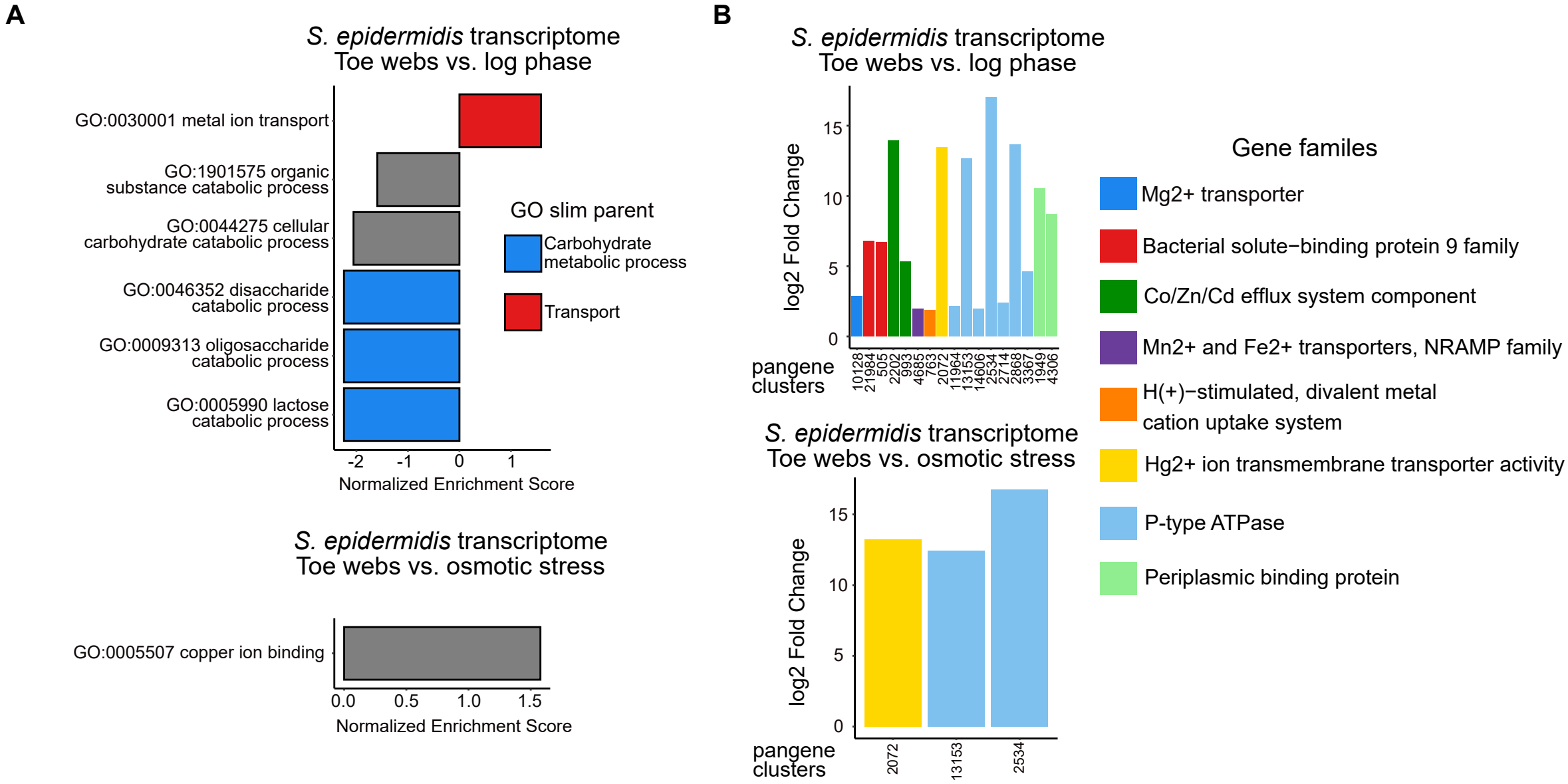
Supplementary Figure 4: Differentially expressed microbial orthologous groups between different skin sites. (A) Volcano plot of expressed bacterial and fungal orthologous groups (OGs) in cheeks (n=22) relative to volar forearms (n=18). (B) Volcano plot of expressed bacterial orthologous groups (OGs) in toe webs (n=18) relative to volar forearms (n=18). (C) Pathway map of heme biosynthesis from glutamate in prokaryotes. Enzymes highlighted in red were ≥2-fold upregulated in toe webs relative to volar forearm (adjusted p-value<0.05)



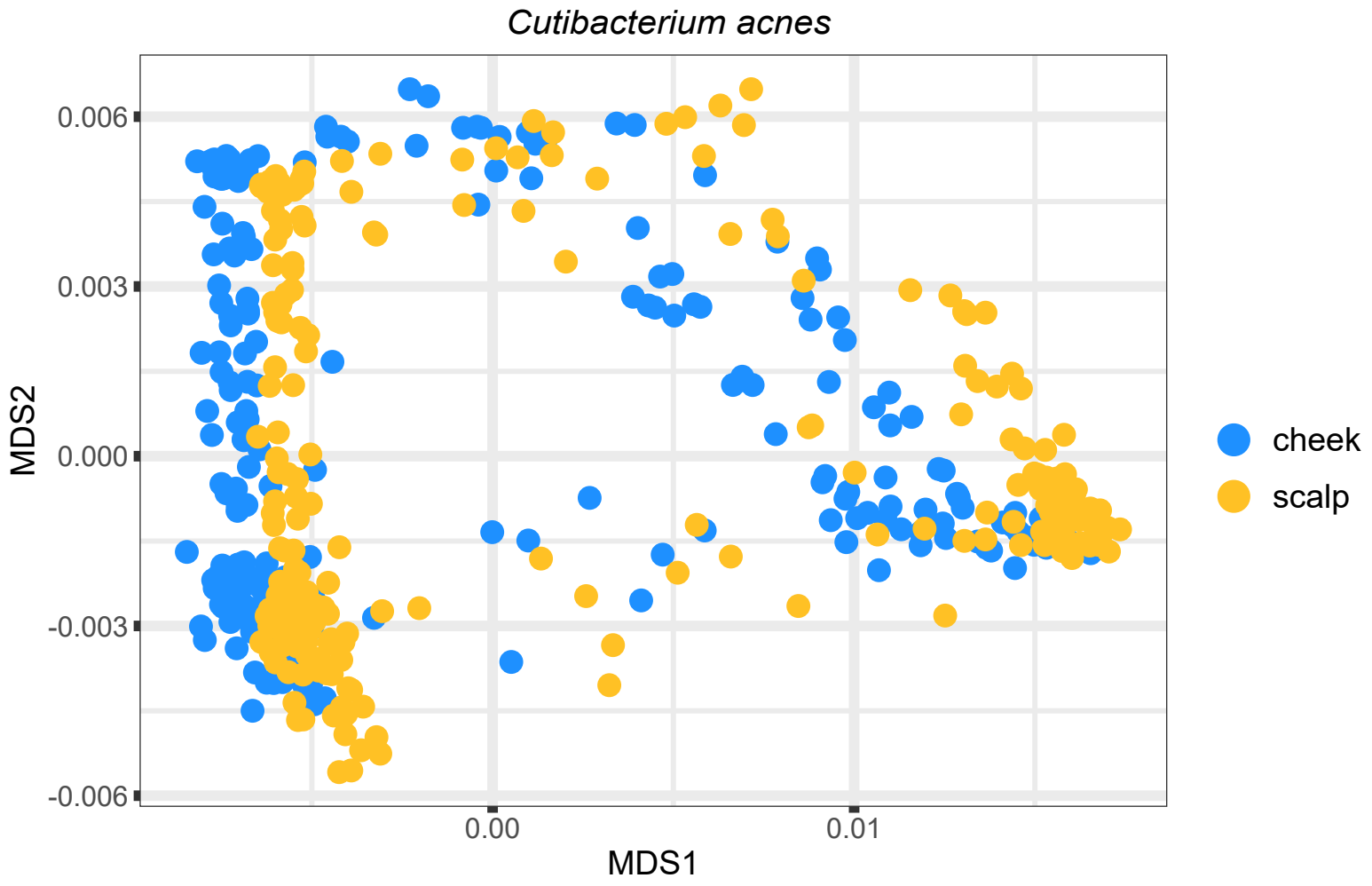
Supplementary Figure 5: Comparison of *in vivo* and *in vitro* gene expression in Staphylococcus epidermidis (A) Principal component analysis (PCA) plot for *S. epidermidis* gene expression profiles from various *in vitro* and *in vivo* samples from the same experimental batch (this study only). (B) Principal component analysis (PCA) plot for *S. epidermidis* gene expression profiles from various *in vitro* conditions, keeping the same linear transformation as the PCA plot in main figure 3E. The different shapes represent three different experimental batches of *in vitro* samples. The y-axis is broken between -20 and -40 for visual clarity.



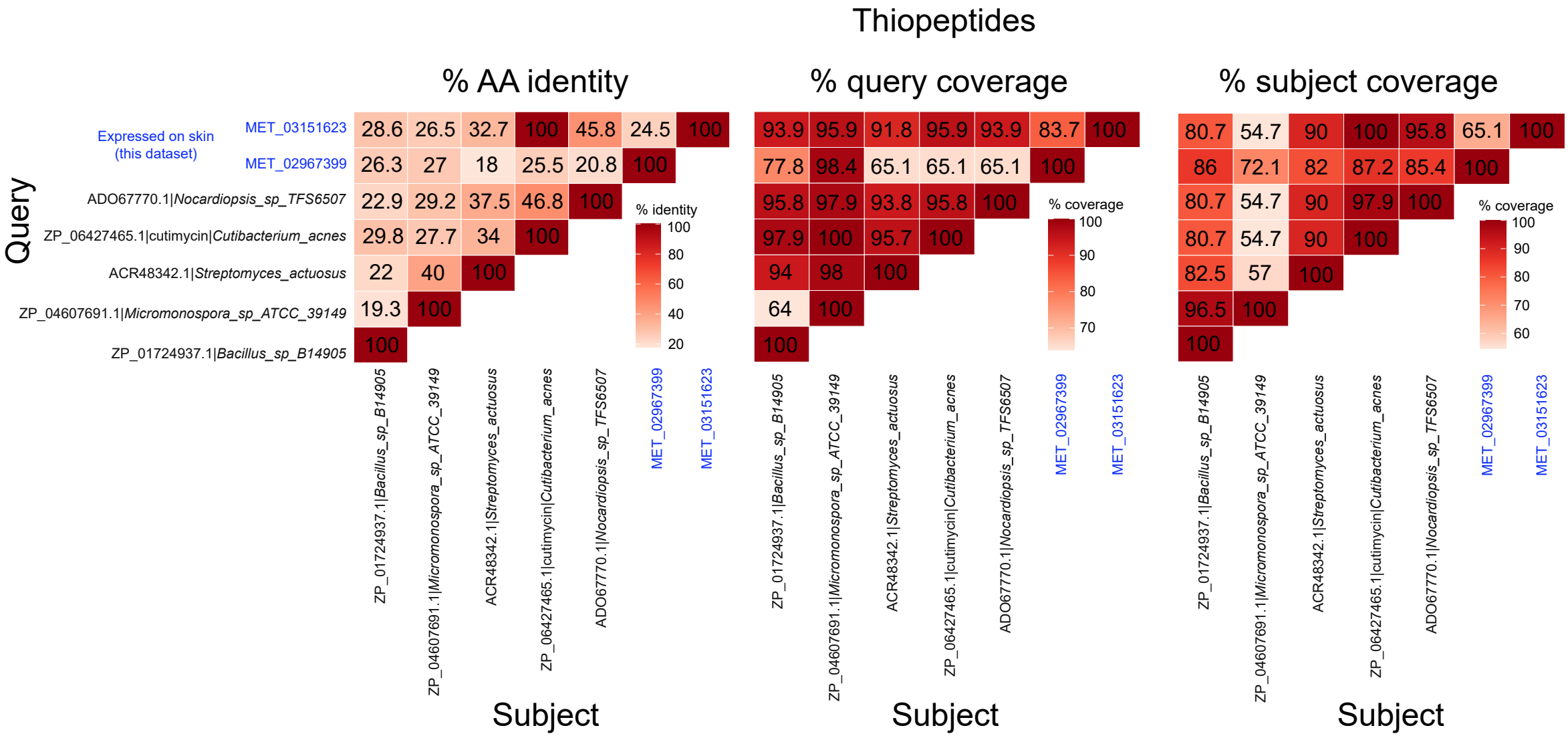
Supplementary Figure 6: Enriched gene sets in *Staphylococcus epidermidis* transcriptomes for sebaceous sites versus *in vitro* conditions. Barplots of normalized enrichment scores derived from gene set enrichment analysis of differentially expressed genes between sebaceous (scalp and cheek) sites (n=6) and various *in vitro* conditions (n=9, 6, 9 for log phase, osmotic stress & stationary phase respectively). Only gene sets or pathways which were significantly enriched in all 3 comparisons are displayed here. Libraries representing sebaceous sites had ≥200,000 *S. epidermidis* reads each.



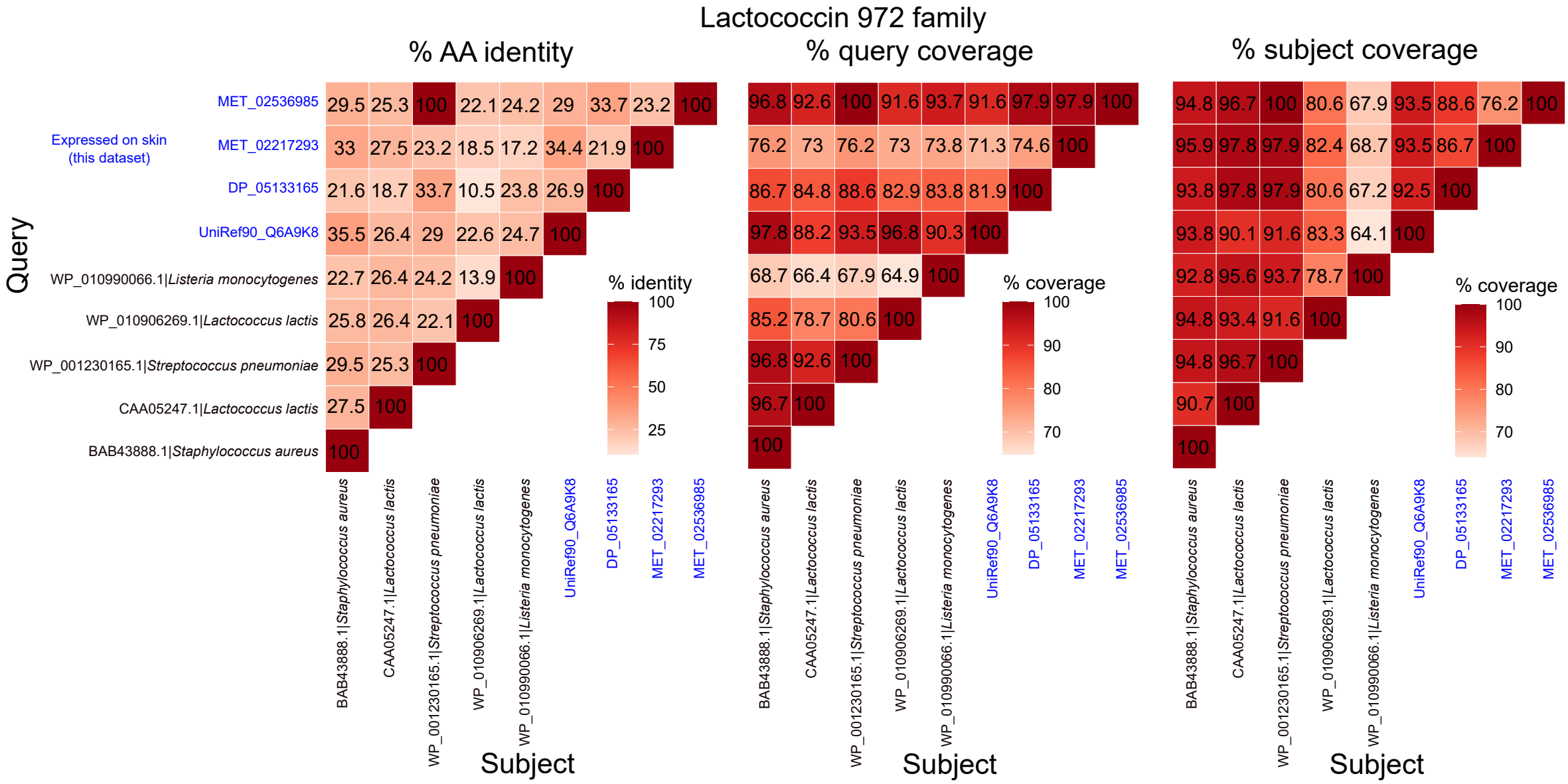
Supplementary Figure 7: Enriched gene sets in *Staphylococcus epidermidis* transcriptomes for toe webs versus *in vitro* conditions. (A) Barplots of normalized enrichment scores from gene set enrichment analysis between toe web sites (n=12) and various *in vitro* conditions (n=9, 6, 9 for log phase, osmotic stress & stationary phase respectively). (B) Barplots showing log2 fold change of genes related to metal ion transport (top) or copper ion binding (bottom), relative to *in vitro* conditions. Libraries representing toe webs had ≥200,000 *S. epidermidis* reads each.



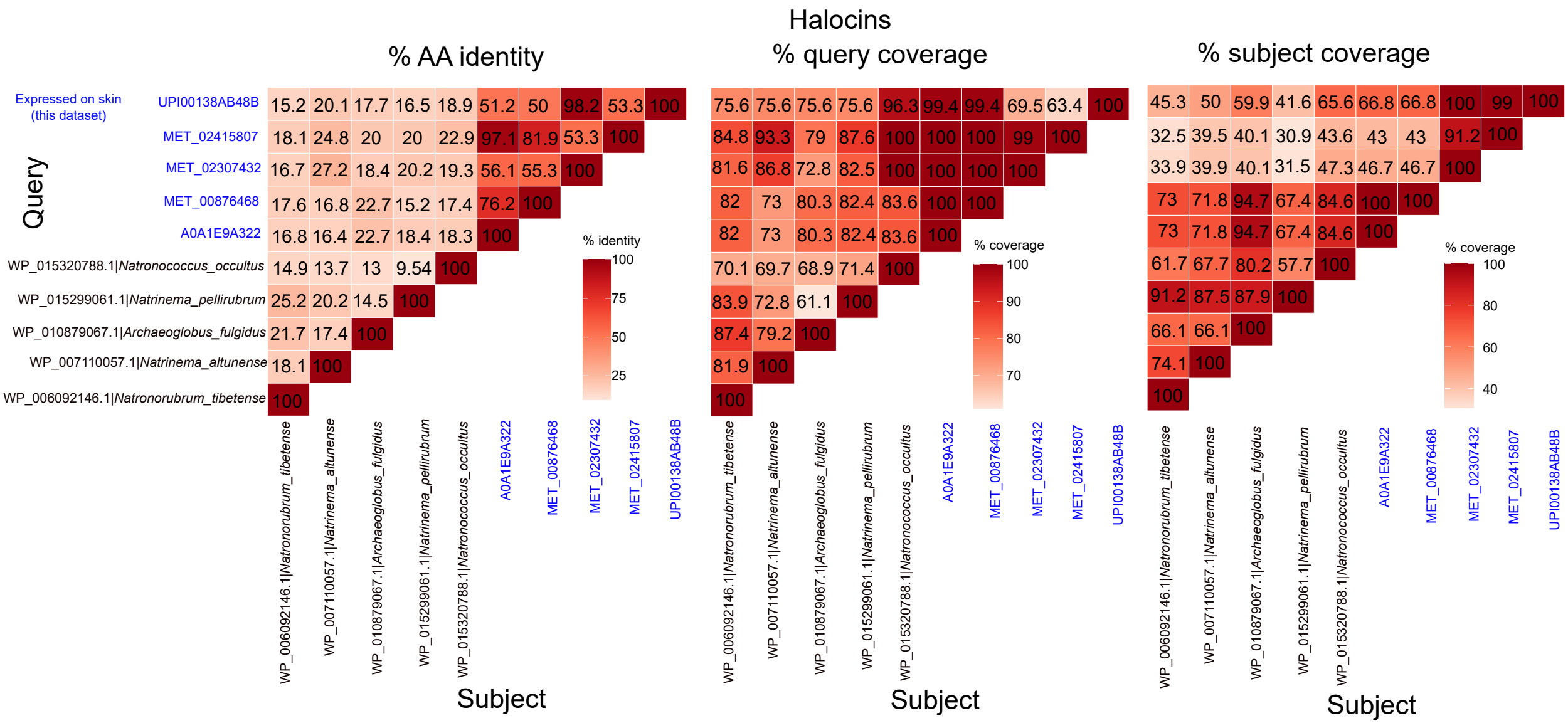
Supplementary Figure 8: Comparison of genome-wide metabolic fluxes for *C. acnes* between cheek and scalp sites. MDS plot showing variation in metabolic fluxes between two *in vivo* conditions: scalp (n=19) and cheeks (n=21) for *C. acnes.* Libraries for *in vivo* conditions had ≥200,000 *C. acnes* reads each.



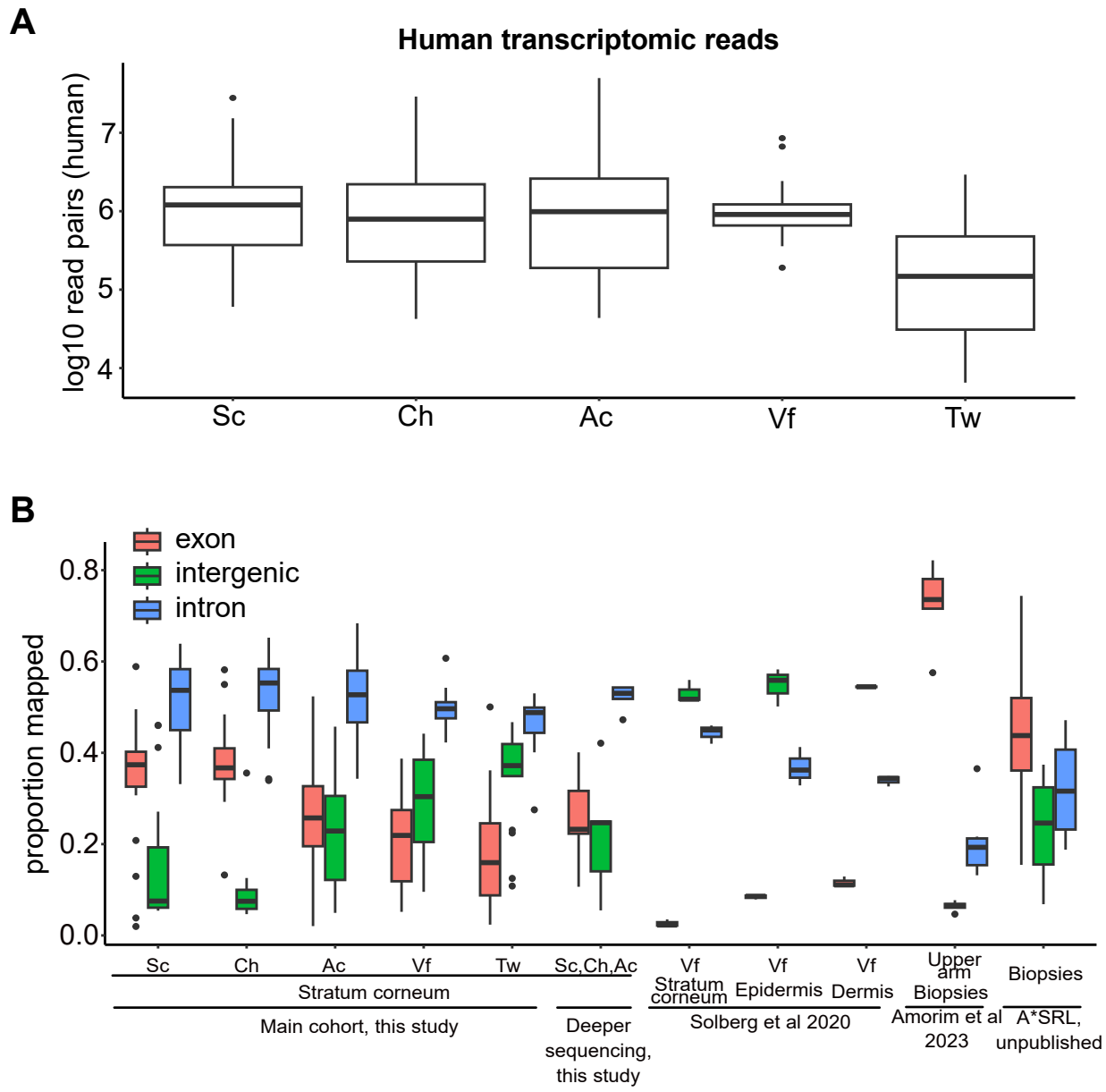
Supplementary Figure 9: Percentage amino acid identity and coverage statistics from pairwise comparisons of thiopeptides. Heatmaps showing various pairwise comparisons of thiopeptide amino acid sequences. Thiopeptide sequences were identified using profile hidden Markov models. Peptides highlighted in blue are expressed on skin in this dataset. Other peptides were chosen representatives from existing databases.



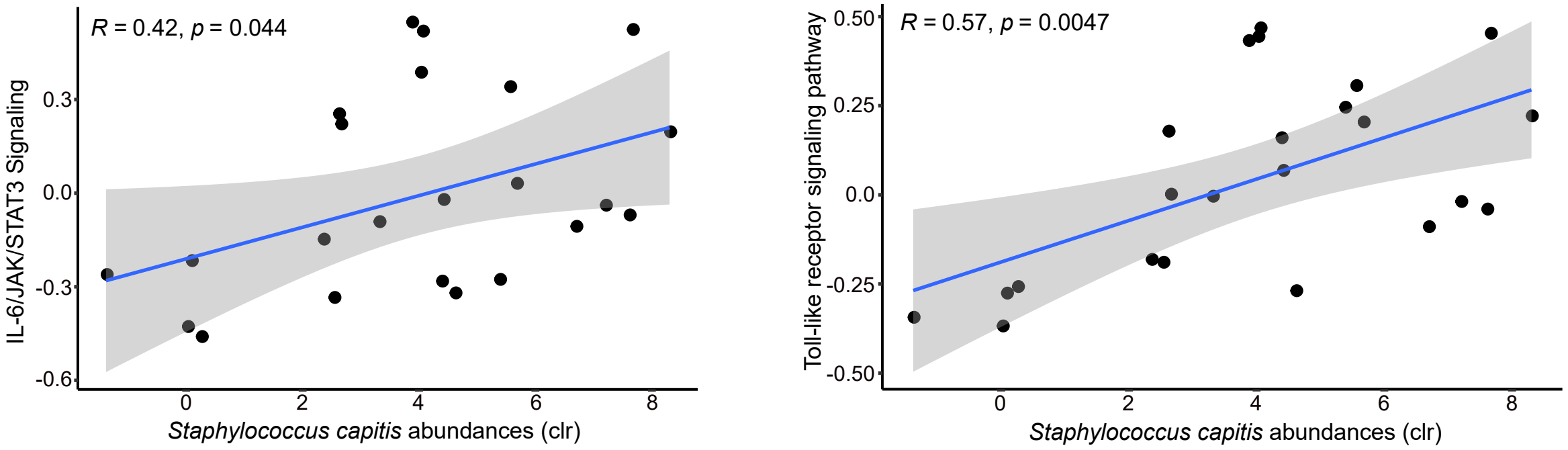
Supplementary Figure 10: Percentage amino acid identity and coverage statistics from pairwise comparisons of bacteriocins of the lactococcin 972 family. Heatmaps showing various pairwise comparisons of peptide sequences. Peptides of the lactococcin 972 family were identified using profile hidden Markov models. Peptides highlighted in blue are expressed on skin in this dataset. Other peptides were chosen representatives from existing databases.



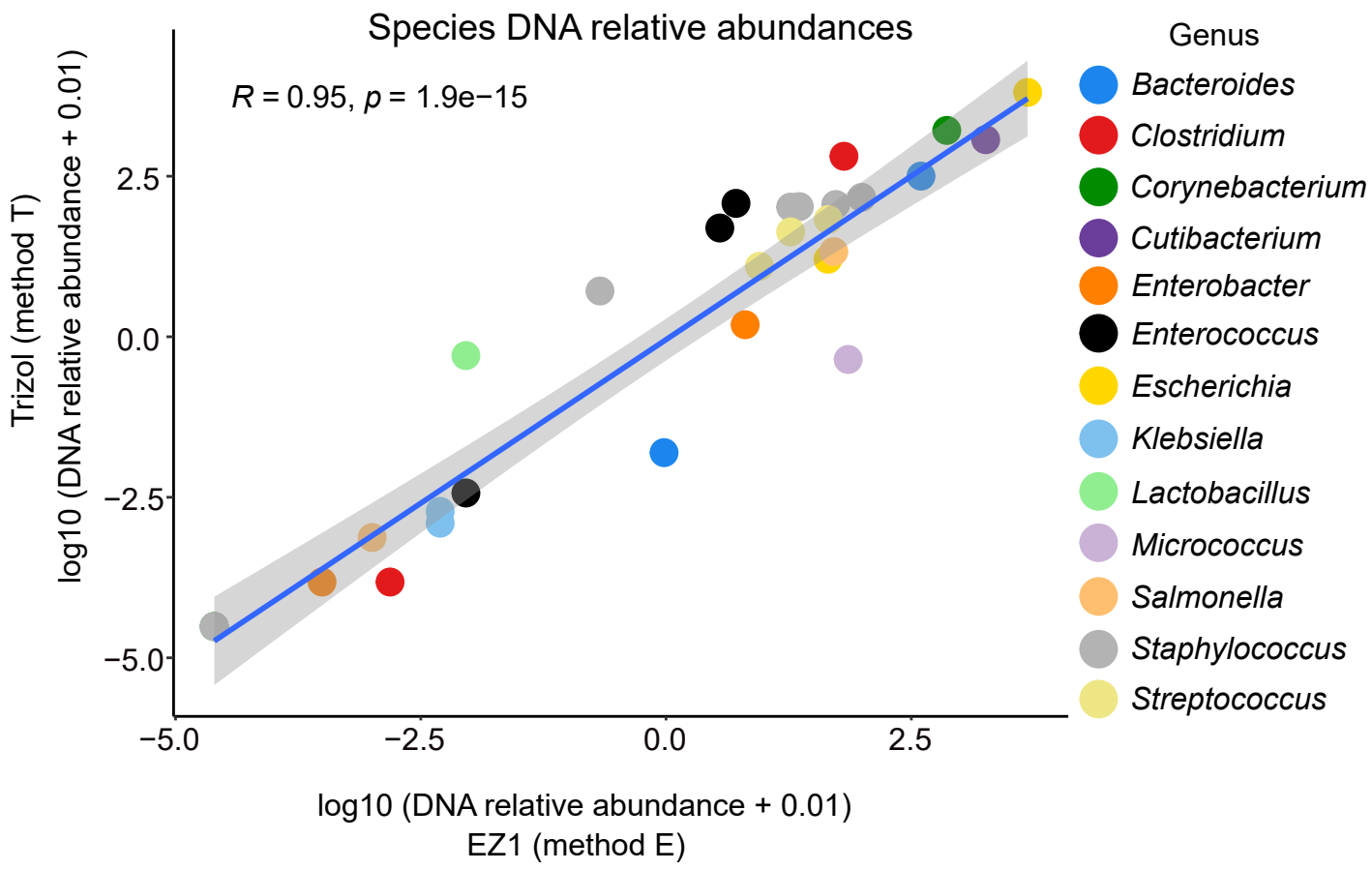
Supplementary Figure 11: Percentage amino acid identity and coverage statistics from pairwise comparisons of halocins. Heatmaps showing various pairwise comparisons of peptide sequences. Peptides of the halocin family were identified using profile hidden Markov models. Peptides highlighted in blue are expressed on skin in this dataset. Other peptides were chosen representatives from existing databases.



Supplementary Figure 12: Quality control metrics for human reads in skin metatranscriptomic libraries. (A) Boxplots depicting number of sequenced human reads across different skin sites from this study. (B) Boxplots showing the proportion of human RNA reads mapping to exons, intergenic regions and introns. The full cohort for this study comprised of n=24, 23, 19, 18, 18 libraries for scalp (Sc), cheek (Ch), antecubital fossa (Ac), volar forearm (Vf) and toe web (Tw) respectively. The samples in this study subjected to deeper re-sequencing comprised of n=2, 3, 3 libraries for Sc, Ch and Ac respectively. The samples for Solberg et al 2020 comprised n=3 libraries each for the stratum corneum, epidermidis and dermis. The samples for Amorim et al comprised n=6 libraries for upper arm biopsies. The biopsies from A*SRL comprised n=4 libraries.



Supplementary Figure 13: Correlations between estimated human pathway activity and *Staphylococcus capitis* abundances. Scatterplots showing positive correlations between immune pathway activity estimated by gene set variation analysis (GSVA) and centered log ratio transformed abundances of *Staphylococcus capitis* on cheeks.



Supplementary figure 14: Correlations between microbial species-level DNA abundances from mock communities extracted using two different methods. Scatterplot comparing relative abundance profiles for microbial DNA extracted with a trizol method (method T) versus DNA extracted using a standard Qiagen kit (method E). Results are from three different mock communities of common human microbes. Taxonomic profiles were obtained using Kraken 2 with default parameters. The Pearson correlation coefficient and corresponding p-value are shown in the figure.

Supplementary Methods

Kitome removal parameters

Potential reagent and laboratory contamination-associated species (the “kitome”) were identified and removed via a multi-step process (**Supplementary Figure 1A**). Swab extraction controls (n=7 negative handling controls) were sequenced from fresh swabs unexposed to human skin. Species with ≥0.1% relative abundance in any negative control constituted an initial list of 70 candidate genera. The main taxa found in the negative controls consisted of environmental microbes such as species belonging to *Janibacter, Elizabethkingia, Achromobacter* and *Bradyrhizobium*. Some of these negative controls also contained signals from microbes commonly found on skin such as *Cutibacterium acnes, Moraxella osloensis* and *Malassezia* species. Each of these 70 genera were classified as kitome candidates if they were not previously reported on skin or in skin diseases, based on hits to the Disbiome database[9], the MicrophenoDB[10] or a PubMed literature search (last accessed 24th Nov 2022) using the search terms ([Genus name] AND "Skin" AND "microbiome" AND "human"). Pairwise Pearson correlations between all species in the metagenomes or metatranscriptomes and these kitome candidates were calculated using SparCC[11] (**Supplementary Data 12**). Any species with a strong positive correlation ($r \geq 0.8$) with kitome candidate species were also marked as potential kitome contaminants and removed from downstream analyses. We observed that abundances of previously reported[12] environmental and kit contaminants such as species of the genera *Achromobacter, Bradyrhizobium, Mycolibacterium, Mycobacterium* and *Brevundimonas* were not strongly positively correlated with skin or oral microbes and could hence be easily distinguished from them for removal (**Supplementary Figure 1B**). Species level taxonomic abundances of sequenced

negative controls and the list of contaminant genera and species identified in this manuscript can be found in **Supplementary Data 1**. Note that the signals from these handling controls would vary across laboratories and are not exhaustive.

Minimizer thresholds for metatranscriptomic false positives

False positive species assignments for metatranscriptomic reads were identified using empirically determined minimizer thresholds. Specifically, RNAs from three biological repeats (MHS445, MHS589 and MHS590), each comprising a mix of non-skin bacteria *Plesiomonas shigelloides*, *Vibrio vulnificus* and *Listeria monocytogenes* were extracted, processed, and sequenced in the same way as other metatranscriptomes in this study. Random subsamples of these libraries were added to a skin metatranscriptome (MHS413) such that read pairs from *P. shigelloides*, *V. vulnificus* and *L. monocytogenes* were each present at relative abundances of approximately 0.1%, 1% or 10% of all reads in the metatranscriptome. These metatranscriptome mixtures were subsequently rarefied to either 10^4 or 10^6 read pairs and reads were classified using Kraken2. For each true positive spike-in species (*P. shigelloides, V. vulnificus and L. monocytogenes*), false positive classifications of reads to other species belonging to the genera *Listeria*, *Plesiomonas*, *Vibrio*, *Pseudomonas*, *Salmonella* and *Klebsiella* were identified. The $\log_{10}$ ratio of unique species-specific minimizers per million microbial reads was plotted against $\log_{10}$ paired counts of each species to determine that 10^4 unique minimizers per million microbial reads could distinguish most true positive species assignments from false positives (**Supplementary Figure 1C**). Species assignment of metatranscriptomic reads were considered true positives if there were ≥10^4 unique Kraken2 minimizers per 1 million microbial reads or had ≥10 read pairs for the species with ≥10× more unique minimizers than read pairs. Species contributing

<0.1% RNA relative abundance were removed from metatranscriptomes to avoid false positives from incorrect read classification. Species contributing <0.1% DNA relative abundance were also removed from each metagenome unless they were also present in the matched metatranscriptome (**Supplementary Figure 1C**).

Mock communities

For metagenomic spike-ins, a mock community of 3 different bacteria (*Vibrio vulnificus* ATCC® 29307™, *Plesiomonas shigelloides* ATCC® 51903™, and *Listeria monocytogenes* ATCC® 35152™) was created by culturing each bacterial species in their respective growth media as recommended by ATCC. A growth curve was established for each bacterial species and the optical density (OD) of each culture was used to estimate the number of colony forming units (CFUs). Equal number of CFUs for each species were resuspended in DNA/RNA shield (Zymo Research; Cat. No. ZYR.R1100) and pooled together to create a mock community stock of concentration 1×10^5 CFU/µL. Mock community stocks were aliquoted and stored at -80°C. Other mock communities for quality control testing were prepared to consist of 10 different bacterial species commonly part of human skin and gut microbiomes (**Supplementary Table 2**). Each species was cultured according to the recommended growth conditions and was counted using a hemocytometer. Cultures were centrifuged and each cell pellet was re-suspended in 100-500µL of DNA/RNA Shield (Zymo Research; R1100). Samples were stored at -80°C prior to combining the cells according to the abundances summarized in **Supplementary Table 2**.

Extraction quality control

To assess the lysis efficiencies of different protocols, each mock community (**Supplementary Table 2**) was extracted with EZ1 DNA Tissue Kit (Qiagen; 953034;

EZ1 method) and a separate aliquot with TRI Reagent® (Zymo Research; Cat. No. R2050-1-200; Direct-zol method). DNA was obtained from the Direct-zol method using additional steps after lysis according to the TRIzol™ Reagent (DNA isolation) User Guide. DNA was cleaned up and concentrated with 2× AMPure XP beads (Beckman Coulter; A63882). DNA was eluted in 30µL of Buffer EB (Qiagen; 19086). Qubit RNA High Sensitivity assay kit (Thermo Fisher Scientific; Q32852) was used to check for RNA contamination. Samples extracted by the Direct-Zol method that had detectable levels of RNA were treated with RNase A (100mg/mL; Qiagen; 19101), cleaned up with 2× AMPure XP beads, and eluted in 30µL of Buffer EB. DNA was used for library preparation and sequenced on a Illumina Novaseq X (2×150 bp reads). This data was used to confirm that metagenomic relative abundances obtained based on DNA extracted from both methods were highly consistent across the different mock communities that were tested (Pearson's R=0.95; **Supplementary Figure 14**).