A multi-omics strategy for the study of microbial metabolism: application to the human skin's microbiome

Louis-Félix Nothias[1,2], Robin Schmid[1,2,3], Allison Garlet[4], Hunter Cameron[5], Sabrina Leoty-Okombi[6], Valérie André-Frei[6], Regine Fuchs[7], Pieter C. Dorrestein[1,2] , Philipp Ternes[7]

1. Collaborative Mass Spectrometry Innovation Center, University of California San Diego, La Jolla, CA 92093, USA
2. Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla CA 92093, USA
3. Institute of Organic Chemistry and Biochemistry of the Czech Academy of Sciences, Prague, 160 00, Czech Republic
4. BASF Corporation, Tarrytown NY 10591, USA
5. BASF Corporation, Research Triangle Park NC 27709, USA
6. BASF Beauty Care Solutions France S.A.S, 69366 Lyon Cedex 07, France
7. BASF Metabolome Solutions GmbH, 10589 Berlin, Germany

ABSTRACT

The revolution of 'omics' technologies highlighted that associated microorganisms (also called microbiota) are integrated into the metabolic functions of their hosts. Yet when performing any particular type of 'omics' experiment, be it metabolomics, transcriptomics, or (meta)genomics, it is extremely difficult to interpret the observed relationships between metabolites, transcripts, and microbial species. This is due to the massive amount of data generated for each 'omics' technology, but also the cognitive challenge of interconnecting these observations and contextualizing them in their biological (eco)system. For these reasons, there is a need for testing methods that can facilitate the translation of these 'omics' experimental observations into putative molecular processes or biological interactions. To accelerate the interpretation of 'omics' data from a description of microbial, transcript, or metabolite identities or abundances into a functional understanding of the interplay between the individual entities of the biological system, we designed a novel multi-omics strategy for the annotation and integration of metabolomics and metagenomics data. We generated metabolome and microbiome datasets by LC-MS/MS based metabolomics profiling and metagenomic sequencing, respectively. Comprehensive metabolite annotations were obtained by molecular networking and computational annotation of fragmentation spectra. Associations between microbes and GNPS molecular networks were predicted by machine learning and visualized as an extensively annotated, nested interaction network in Cytoscape. As a proof of concept, we applied this strategy to scalp swabs from a cohort of healthy volunteers with varying scalp sebum levels and were able to elucidate the antagonistic interaction between two well-characterized microbes, *Staphylococcus epidermidis* and *Cutibacterium acnes*.

INTRODUCTION

The revolution of “omics" technologies highlighted that associated microorganisms (also called microbiota) are integrated into the metabolic functions of living organisms. Their association, involving either mutualism or parasitism, is called a holobiont. An understanding of the molecular basis of the biological processes implicated in holobionts could be translated into biomimetic clinical applications (1). For example, evidence is accumulating for a central role of skin-associated microbiota in the pathogenesis and the prevention of many skin diseases (2), and this microbial diversity offers a reservoir for promoting skin homeostasis, as shown by the use of a strain of *Staphyloccocus epidermidis* in preventing skin neoplasia (3).

Advancements in microbiome sequencing techniques allow the profiling of microbial species and their genomes directly from a given biological or environmental sample, effectively going beyond the limitation of lab-based cultivation techniques. When shotgun metagenomic sequencing is performed, it is also possible to establish the metabolic pathways that are overrepresented, including in a specific microorganism. Yet, the expression of the genetic potential of microorganisms in a holobiont is influenced by many factors including microbial population dynamics, the host genome, and environmental factors. These molecular interactions could be monitored by transcriptomics, proteomics, and metabolomics analysis. In particular, mass spectrometry (MS)-based metabolomics has matured enough to allow the observation of a large scope of metabolites, originating both from microbial and/or host metabolic processes. These metabolic cross-talks are called meta-metabolism. This comprises anabolic and catabolic pathways that govern the cellular functions, but also metabolic processes that mediate interspecies synergistic/antagonistic interactions. In addition, metabolomics complements other omics technologies by highlighting small molecules potentially resulting from complex metabolic processes, either biotic or abiotic, that could not be accessed with other omics technologies. Furthermore, modern MS-based metabolomics facilitates the analysis of thousands of metabolites by collecting their fragmentation spectra (MS/MS) in high resolution and a data-dependent manner, which can be used for metabolite annotation (4).

However, the annotation (putative identification) of metabolites from their MS/MS spectra has for a long time been challenging due to the limited number of reference spectra available in spectral libraries (typically offering less than 5-10% of annotation rate) (5). A further limitation has been the lack of systematic and reliable bioinformatics approaches to characterize many of the metabolites observed. In addition, the set of metabolites efficiently detected by mass spectrometry depends on the chromatographic conditions and ionization methods employed.

One more challenge is the integration and interpretation of data generated by multiple omics technologies. Machine learning models were introduced to generate biological hypotheses for microbe-metabolite interactions (6). Yet, these predicted microbe-metabolite interactions remain challenging to visualize and interpret.

We aimed to devise a comprehensive strategy for the study of the meta-metabolism in holobionts. For this, we have designed a multi-omics (MO) based-methodology to explore relationships between metabolites and microorganisms. Metabolomics analysis is performed in multiple complementary mass spectrometry conditions to facilitate the discovery of diverse metabolites, and their annotation is made possible thanks to the latest advances in computational mass spectrometry methods (7), including systematic class annotation with SIRIUS/CANOPUS (8). Microbial profiling is performed by metagenomic sequencing, and the MO integration is achieved by predicting conditional probabilities of microbe-metabolite

associations. Finally, the results of interactions are visualized as a nested molecular/microbial network that facilitates the analysis of the microbial and host-microbe interactions.

The MO strategy was applied to the study of the metabolomics and microbiome profile of scalp samples from human volunteers with either normal or oily scalp to understand the composition and function of different microbiome profiles on different scalp environments. The combination of GNPS spectral library searches (9) and SIRIUS (8) *in silico* annotations yielded a highly annotated metabolomics dataset that was correlated with physical parameters of the scalp (oiliness and moisture), describing the “scenery” on which the microbial players interact with each other and with the human host. The metabolomics–microbiome data integration then allowed us to take the study one step further towards watching the biological “theater play” that unfolds between the different players that constitute the holobiont.

RESULTS AND DISCUSSION

The multi-omics strategy comprises five components (Figure 1): 1) mass spectrometry-based metabolomics, 2) microbiome profiling, 3) statistical analysis, 4) MO integration, and 5) visualization and interpretation of results. Notably, this MO strategy can be applied to various biological or environmental samples without prior molecular knowledge or hypothesis, as it allows the monitoring of various microorganisms and metabolites. This is made possible by the dual use of metagenomic sequencing on one hand and untargeted mass spectrometry with novel de *novo* annotation capabilities on the other hand. Importantly, the capability to visually navigate the nested networks of predicted interactions between metabolite-microbes pairs offers a powerful integrated framework for exploring the relationships and the nature of the molecules involved.

Mass spectrometry-based metabolomics. Multiple untargeted analyses were performed to cover a wide chemical space of metabolites differing between normal and oily scalps. First, the metabolites were sampled on 74 healthy volunteers (normal scalp, n = 33; oily scalp, n = 41; see Method Section for details) using ethanol-pre-soaked cotton swabs, as previously described (10). To cover a broad range of metabolites, each sample was analyzed by LC-MS under four experimental conditions: Two chromatographic methods (reversed phase [RP18] and hydrophilic interaction [HILIC] chromatography) and two ionization modes (positive and negative electrospray ionization [ESI]).

The data from each of these conditions were processed separately with MZmine (11), using the Ion Identity Molecular Networking (IIMN) workflow for improved identification of ion-types (e.g., adducts and in source fragments) by feature shape correlation analysis (12). Briefly, ion features (m/z, retention time, isotope pattern, MS/MS) are detected and aligned across all samples, grouped with their corresponding MS/MS spectra, and exported for SIRIUS and the IIMN workflow on the GNPS web-platform (9). The fragmentation spectra were then annotated by spectral library search against the reference spectra from the GNPS open community libraries, before molecular networking by a pair-wise modification aware spectral similarity of all MS/MS (13). Optionally, analog compounds were annotated by modification-aware library search.

To expand the annotation coverage, the spectral data from all the metabolites detected were further annotated with SIRIUS (4), providing putative annotations for : A) molecular formulas (14), B) putative structures (15), and C) chemical class (8). The results of

GNPS and SIRIUS were combined and visualized as annotated molecular networks. The annotation of the molecular networks was further enriched with statistical analysis results by displaying the correlation of feature intensity (chromatographic peak area, log scale) with the oiliness of the skin (Sebumeter value).

Metabolite annotations from spectral libraries and molecular networking. To examine the annotation coverage obtained by spectral library matching with the GNPS and the NIST spectral libraries, the sets of annotated features and their distribution in datasets were visualized as UpSet plot (Figure 2A). HILIC provided the largest number of spectral matches (286) in positive ionization mode (HILIC+), but only 84 in negative ionization mode (HILIC−). Reverse phase chromatography gave 105 matches in positive ionization mode (RP18+) and 21 in negative ionization mode (RP18−). The highest overlap of spectral library annotations was observed between the HILIC+ and RP18+ (78) and between the HILIC+ and HILIC− (57). Overall, these results illustrate that each metabolomics dataset contributes to a different metabolic space, depending on the metabolites' polarity and ionization preference.

When comparing the number of annotations to the total number of features, the annotation rate was 2–4 % in each dataset (Figure 2B, Supplementary Figure S1 and Supplementary Table S1). This proportion was increased to 9–10 % by performing spectral library matching in analogue mode, which tolerates a spectral shift for the fragment and precursor ions between the experimental and the library spectra. Yet, this analogue mode is known to have a higher chance of providing incorrect matches. For molecular networking results, the number of nodes in networks was lower for the HILIC+ (16%) and HILIC− (27%) datasets and higher for the C18+ (29%) and C18− (53%). These differences are likely due to different fragmentation behavior for chemical classes preferentially observed in each dataset.

Expanding metabolite annotations with SIRIUS. To increase the annotation coverage, we computationally annotated molecular formulas (14), putative structures (15), and chemical classes (8) using the SIRIUS tools. As observed for the spectral library annotation and molecular networking results, the proportion of annotations with SIRIUS varied per dataset (Fig. 2B, and Supplementary Table S1). Nevertheless, the annotation rate was considerably expanded for all datasets, with a 23.5-fold increase in annotation coverage for the C18+, a 45-fold increase for C18−, a 7.6-fold increase for HILIC+ and a 13.5-fold increase for HILIC−. Chemical class predictions were obtained with CANOPUS for 26% and 29% of the features in the C18+ and C18− datasets, respectively. The rate was lower in the HILIC datasets with 16% and 5% for the HILIC+ and HILIC−, respectively. The source of this difference is likely due to the machine-learning nature of the class prediction method. Indeed, the number of reference spectra available in public spectral libraries acquired in negative ionization mode or observed in HILIC is relatively limited. Thus, fewer reference spectra can be used for the training of the CANOPUS neural network and fewer classes can be predicted reliably in these modes. For putative structure prediction with CSI:FingerID, the annotation rate ranged from 11% (HILIC+) to 30% (HILIC−). Overall, compared to spectral library matching-based annotation, CSI:FingerID increased the network annotation rate by 1.75 (C18+ and HILIC-) to 15.5 fold (C18-).

When combining the annotations from spectral library matching (both direct and analogue mode) with the SIRIUS annotations, the total feature annotation rate reached 52%, 90%, 24%, and 54% for the C18+, C18−, HILIC+, and HILIC− datasets, respectively. Furthermore, when combining these annotations with molecular networking, 67% (+22.4%), 95% (+5.3%), 28% (+14.3%), and 57% (+5.3%) of the features were either directly annotated or connected in an annotated molecular network. This illustrates how combining multiple

annotation strategies allows us to better apprehend the chemical diversity observed in metabolomics datasets.

Evaluating the consistency of the SIRIUS predictions. To evaluate the consistency of the data processing and annotation performed on each dataset, we relied on the following methodology: high-quality spectral library matches were considered ground truth for comparing SIRIUS-based molecular formula and class prediction.

Results showed that the predicted molecular formula matched the spectral library annotation in 66% (n= 44), 70% (n = 10), 90% (n =125), and 95% (n = 59) of evaluation pairs for the C18+, C18−, HILIC+ and HILIC− datasets, respectively (Supplementary Figures S2–S5). Inspection of incorrect molecular formula prediction revealed that often an incorrect adduct assignment was involved (in particular sodiated instead of protonated for the positive ionization mode). This is also supported by the higher annotation consistency for the negative ionization. In addition, the lower consistency rate in C18 datasets could also be related to the high number of isobaric lipid species observed in this chromatographic condition, which are challenging to annotate unambiguously even in high-resolution mass spectrometry.

An assessment of the chemical class prediction accuracy was not feasible for the datasets in negative ionization mode, as only 10 pairs of high-confidence spectral library and chemical class prediction were available for each. However, for the positive ionization mode, 44 and 120 pairs were available for C18+ and HILIC+, respectively. The chemical class was correctly predicted for C18+ and HILIC+ datasets, respectively, at the following accuracy: 84% and 92% for the superclass level; for 80% and 90% at the class level; and 57% and 86% for the subclass level. Interestingly, this evaluation shows that the chemical class prediction was relatively robust for the C18+ dataset even if the molecular formula prediction accuracy was comparatively lower at 66%.

Metabolomic differences between normal and oily scalp. We next assessed the correlation of each observed feature's intensity with oily or normal scalp. After a filtering step to remove features that contributed little to biological information (high feature intensities in blank samples or high technical variability), the feature intensities were correlated with scalp oiliness (assessed with a Sebumeter) and scalp moisture (skin conductance measured using a Skicon device) by ANOVA. The statistical analysis comprised 44,570 features from the four datasets combined.

Figure 3 shows selected molecular networks from the C18+ (Figure 3A) and HILIC+ (Figure 3B) datasets. Symbols represent the annotation method of the feature while colors highlight statistically significant differences between skin types. The molecular network in Figure 3A contains several fatty acid methyl and ethyl esters as well as monoacylglycerols that are significantly increased on oily scalp (GNPS spectral library annotation, inverted triangle). Interestingly, nodes with a direct GNPS spectral library annotation consistently show a positive correlation with scalp oiliness while most other nodes do not.

Fatty acid methyl esters and their sulfonated or ethoxylated derivatives are common constituents of cosmetic products (16). This may point to different use of cosmetic products by volunteers with oily scalps, e.g., more frequent hair-washing. Although volunteers agreed to refrain from using shampoo or other cosmetics products starting two days before the first sample collection, some cosmetics ingredients were detected on the skin showing that they perdure on the skin for several days (10).

Monoacylglycerols may originate from triacylglycerols, a major constituent of sebum, by the action of microbial lipases. Fatty acid ethyl esters are established products of microbial metabolism (17). Most of the available literature on bacteria and yeasts capable of

producing fatty acid ethyl esters is focused on their biotechnological utilization for making biodiesel. We are not aware of any studies on their physiological role within the ecosystem of the human scalp. We speculate that in a lipid-rich environment like sebum, fatty acids (liberated from triacylglycerols by bacterial lipases) constitute a nearly unlimited source of carbon and energy. In such an environment, the rate at which acetyl-CoA is being produced by microbial fatty acid *β*-oxidation may exceed its consumption in the TCA cycle. Reduction to ethanol followed by esterification to fatty acids, which are abundant, may constitute a safe way of disposing off excess carbon while utilizing the NADH being produced during *β*-oxidation.

The network in Figure 3B comprises nodes that are significantly decreased in oily scalp. While there are no GNPS annotations, the SIRIUS annotations are consistent with most nodes representing one or several classes of polar lipids containing one or three N in their molecular formula, e.g., lysosphingolipids. Sphingolipids are typical stratum corneum lipids, thus a decrease of sphingolipids or sphingolipid metabolites on oily scalp points to a shift in the relative proportions of sebum-derived and stratum corneum-derived metabolites. Overproduction of sebum by volunteers with oily scalps will dilute stratum corneum-derived metabolites (18), thus changing the environment for the skin microbiota.

Conclusions from the metabolomics data analysis. These examples show that biologically meaningful conclusions can be drawn from the metabolomics data from volunteers with oily or normal scalps. Despite the richness in feature annotation achieved by the combination of different analytical methods and advanced computational annotation tools, the conclusions drawn from the metabolomics data so far are rather descriptive. Mechanistic insight into the functioning of the skin microbiota is either indirect or speculative. We therefore asked whether linking metabolites to specific microorganisms would elevate the data analysis from a descriptive to a more functional level. This question was addressed by two complementary approaches: 1) Systematic mass spectral search for features in the metabolomics datasets that were also detected in publicly available dataset from microbial monocultures using the yet to be published microbeMASST tool, a domain specific MASST tool (19), where one can search MS/MS data against data of ~60,000 taxonomically curated cultures of microbes (https://masst.ucsd.edu/microbemasst/), and 2) MO integration of the metabolomics datasets with a complementary metagenomic sequencing dataset using MMvec (6).

Search for microbial metabolites. MASST (19) is a recent addition to the GNPS ecosystem and allows to search whether a fragmentation spectrum in a metabolomics dataset has been previously observed in any dataset from the public data repository part of the GNPS/MassIVE knowledge base. microbeMASST (https://masst.ucsd.edu/microbemasst/) performs a MASST search against a curated set of metabolomics datasets from microbial monocultures. To systematically link metabolites from the skin swabs to specific microorganisms, we searched all feature's fragmentation spectra from networks with at least three connected nodes with microbeMASST in batch mode.

The network shown in Figure 3A turned out to be special in comprising many features with a positive correlation with scalp oiliness (red fill color) as well as many microbeMASST hits (green labels). Detailed annotations for each hit can be found in Supplementary Figures S6–S26. While some of the features in this network may originate from cosmetic products as hypothesized above, the prevalence of microbeMASST hits suggests that many of the features are in fact of microbial origin. This lends support to the hypothesis that fatty acid ethyl esters are products of the microbial metabolism of fatty acids derived from sebum lipids

(see above). Such microbial metabolites of host-derived molecules can be seen as products of the holobiont's metabolism.

Metabolomics–microbiome data integration. As a complementary approach to identifying links between metabolites and specific microorganisms, we integrated the metabolomics data with microbiome data acquired by metagenomic sequencing of skin swabs taken from the same volunteers at the same time as the metabolomics samples (see Methods Section for details). The MO integration was performed with MMvec, a neural network that learns the co-occurrence probabilities of metabolite/microbe pairs (6).

MMvec predicts the abundance of all metabolites, given the presence of each microbial taxon. In doing so, it goes through the microbial taxa one-by-one, each time considering all metabolite features simultaneously. To decrease the risk of overfitting the neural network due to the large number of metabolite features, we performed data reduction on the metabolomics dataset, taking advantage of the molecular networks generated by GNPS.

Metabolomics data reduction using molecular networks. We observed that the feature intensities within one molecular network were in many cases positively correlated with each other (Supplementary Figure S27), presumably due to biological similarities between structurally related metabolites, but also due to different features representing the same metabolite (e.g., different adducts). We therefore used as input for MMvec the median value of all metabolite features within each molecular network. Considering only networks with at least three connected nodes, the dimensions were reduced to 1,225 MN representing 9,076 features out of 44,570 total features.

The metabolomics–microbiome data integration results were visualized as a nested microbe–metabolite network. Figure 4 shows microbial taxa as circles and molecular networks (each comprising a number of metabolite features) as rounded squares. The edges denote associations between microbes and metabolites: A red edge indicates that the metabolite features within the molecular family (molecular network connected component) were (on average) more abundant when the microbe was present (positive association), and a blue edge indicates that they were less abundant (negative association). In addition, the border color of the rounded squares shows how the metabolite features within each molecular family are (on average) correlated with skin oiliness (Sebumeter value): A red border indicates a positive correlation, a blue border indicates a negative correlation, and a gray bord means no significant correlation.

Of particular interest is the interaction between the microbes *Staphylococcus epidermidis* and *Cutibacterium acnes* and the three molecular families (two from C18− and one from HILIC+) associated with both. These molecular families are positively associated with *S. epidermidis*, but negatively with *C. acnes*, suggesting an antagonistic interaction between *S. epidermidis* and *C. acnes*, in which metabolites represented by these molecular families are involved. In addition, the metabolite features in the three molecular families are (on average) positively correlated with skin oiliness, suggesting that a possible link to the amount of sebum may exist in these interactions.

The role of *S. epidermidis* and *C. acnes* on the human skin, and their antagonistic interaction is well described in the literature (20–22). The available literature suggests that on the molecular level, the antagonistic interaction between *S. epidermidis* and *C. acnes* involves (among other factors) small molecules produced by the microbes themselves or through conversion of host metabolites by the microbes. In addition, the relative abundance of *S. epidermidis* and *C. acnes* may reflect their respective preference for microenvironments with different relative amounts of host-derived molecules.

Conclusions from the metabolomics–microbiome data integration. The metabolomics–microbiome data integration results suggest that the metabolites represented by the three molecular networks play a key role in the antagonistic interaction between *S. epidermidis* and *C. acnes*. Although none of the metabolite features are annotated with a spectral library match, the computational annotations are consistent with the respective metabolites being lipid species containing a single N in their molecular formula. Although the chemical class prediction was not homogenous for these molecular families due to overlapping chemical classes of the *Classyfire* ontology used by SIRIUS, the metabolites may belong to one or more N-containing lipid classes like lysophospholipids (e.g., lyso-PE, lyso-PC, lyse-PS), *N*-acylhomoserine lactones (AHL), ornithine lipids (OL), or ceramides.

To our knowledge, none of these lipid classes has been previously described as being involved in the antagonistic interaction between *S. epidermidis* and *C. acnes*. As indicated by their positive association (red edges in Figure 4), the metabolites represented by all three molecular networks are most likely produced or induced by *S. epidermidis* and may contribute to controlling the abundance of *C. acnes*, with which they are negatively associated (blue edges in Figure 4). Alternatively, their presence may be characteristic of a microenvironment that favors the growth of *S. epidermidis* but disfavors *C. acnes*. Lysophospholipids may result from the action of bacterial lipases on PE, PC, or PS. *S. epidermidis* is known to produce several lipases that hydrolyse host lipids (23). AHLs are well-described quorum-sensing molecules that may coordinate the activity of *S. epidermidis* cells (24, 25). Bacteria commonly produce OL under phosphate-limiting conditions (26). While ceramides are typical stratum corneum lipids (27). A positive or negative association of a microbe with ceramides may indicate a preference for a microenvironment enriched in stratum corneum or sebum lipids, respectively.

Final remarks. Our MO strategy pointed us to key microbial players in skin homeostasis and highlighted their metabolic interactions. microbeMASST added credibility to the hypothesis that the fatty acid ethyl esters in Fig. 3A are a product of microbial metabolism by showing that these and similar compounds have been observed as microbial metabolites in independent studies under different conditions. MMvec finally allowed to go beyond (functional hypotheses built on) metabolite observation towards an empirical investigation of microbial function. Both the observed interaction between *S. epidermidis* and *C. acnes* and the presumed production of fatty acid ethyl esters from sebum lipids likely represent large-scale activities on the human scalp. To reveal more subtle details of the interactions between microbes and their human host, more highly powered studies as well as independent confirmation will be required. Nevertheless, this MO integration allowed the generation of hypotheses on likely meta-metabolism interactions that could govern the equilibrium of the skin ecosystem and could inspire the design of follow-up studies. The MO strategy could be conveniently applied to other studies and holobionts, and we anticipate that such integrative approaches will help to harness the potential of MO studies.

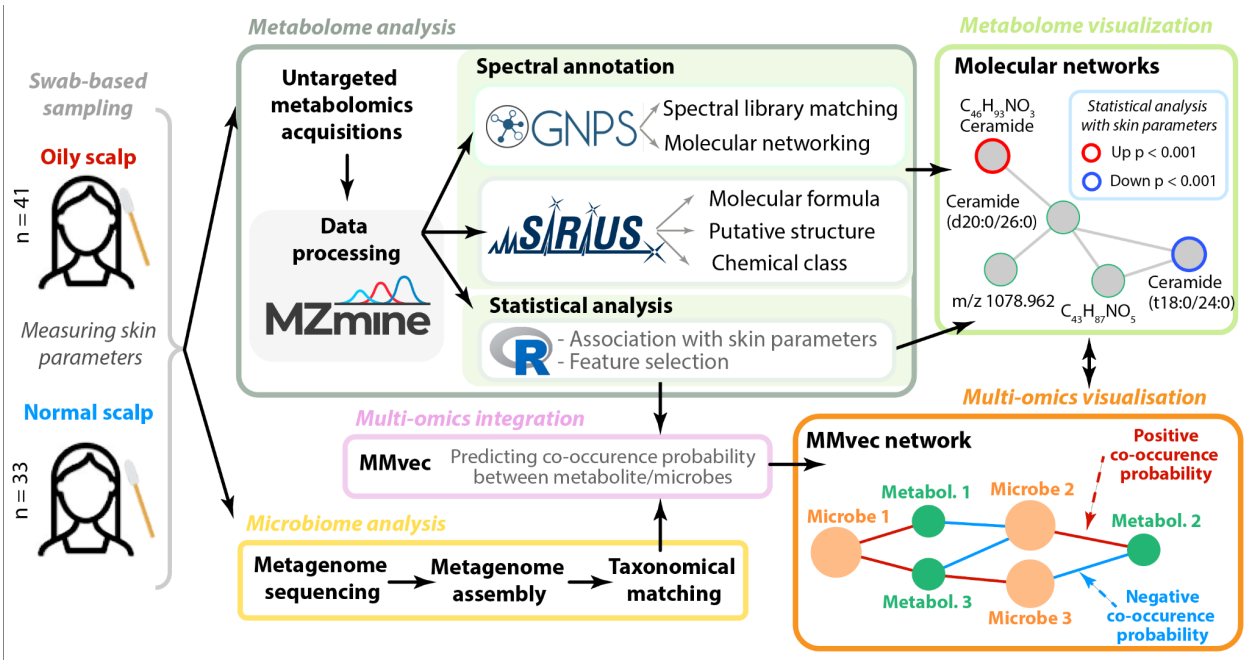


Figure 1. Overview of the study design and multi-omics analysis strategy relying on both untargeted metabolomics and metagenomics sequencing. The omics integration was performed with MMvec and the results visualized as a microbe-metabolite network.

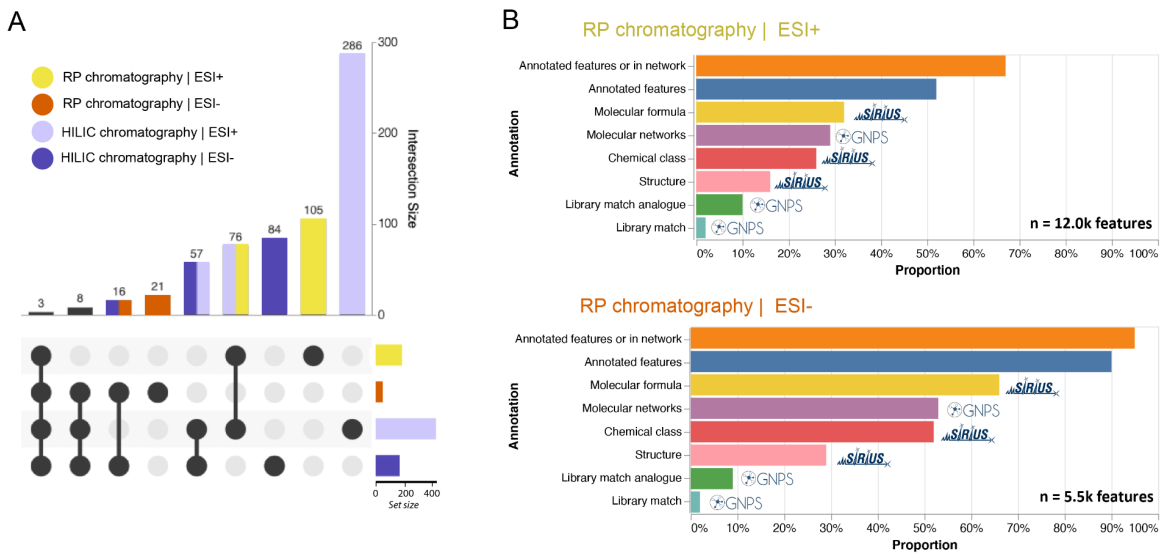


Figure 2. A) Distribution of metabolites annotations per datasets (RP18 = reversed phase, HILIC = hydrophilic liquid interaction, + = positive ionization mode, and − = negative ionization mode) and B) Proportion of LC-MS features annotated by GNPS or SIRIUS in the RP18 datasets (top, positive ionization mode; bottom, negative ionization mode).

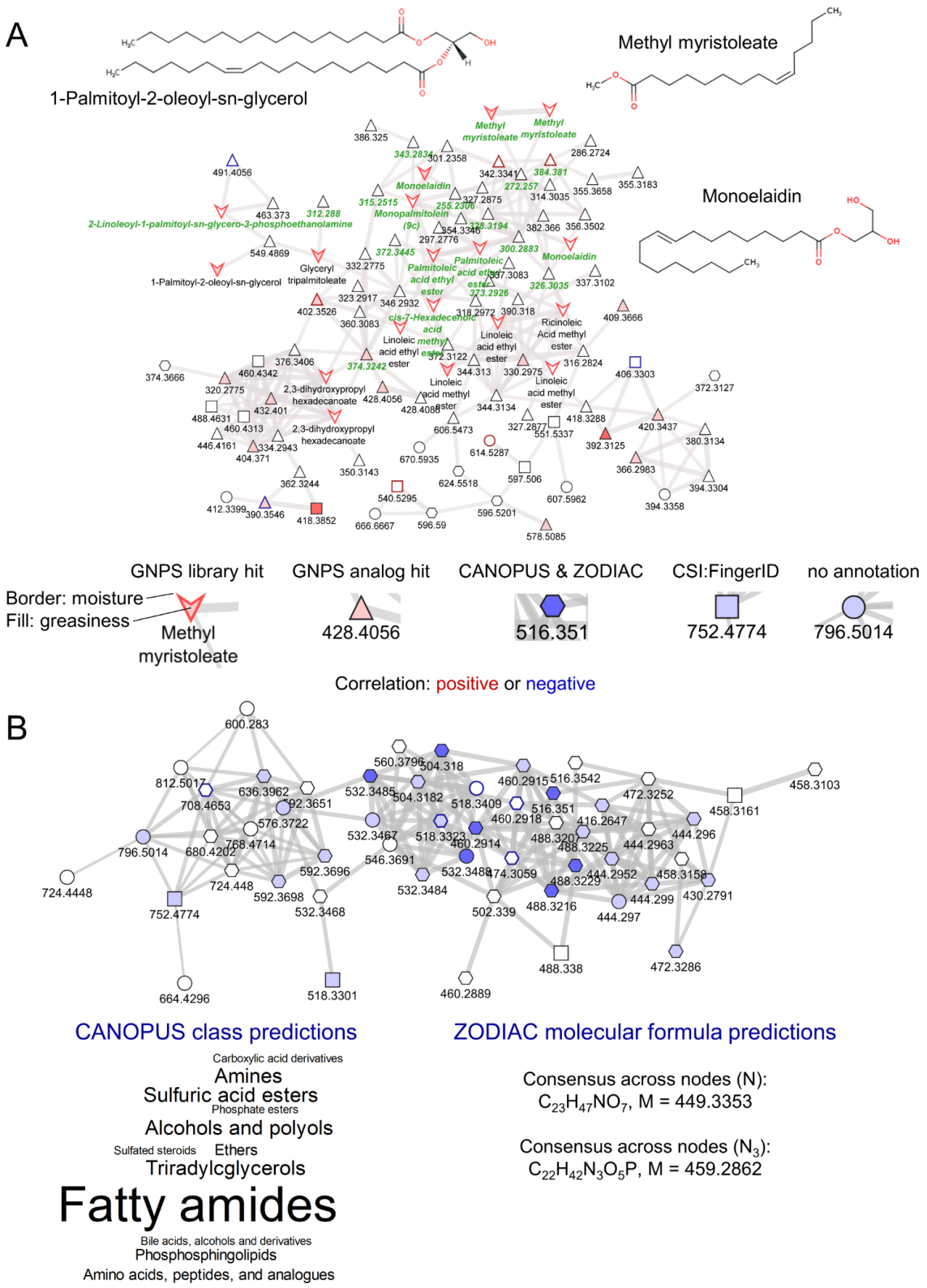


Figure 3. Molecular networking families: A) Example from the C18+ dataset showing structures for selected GNPS library spectral library matches. Green labels indicate a microbeMASST spectral match to a microbial monoculture. B) Example from the HILIC+ dataset showing consensus SIRIUS annotations.

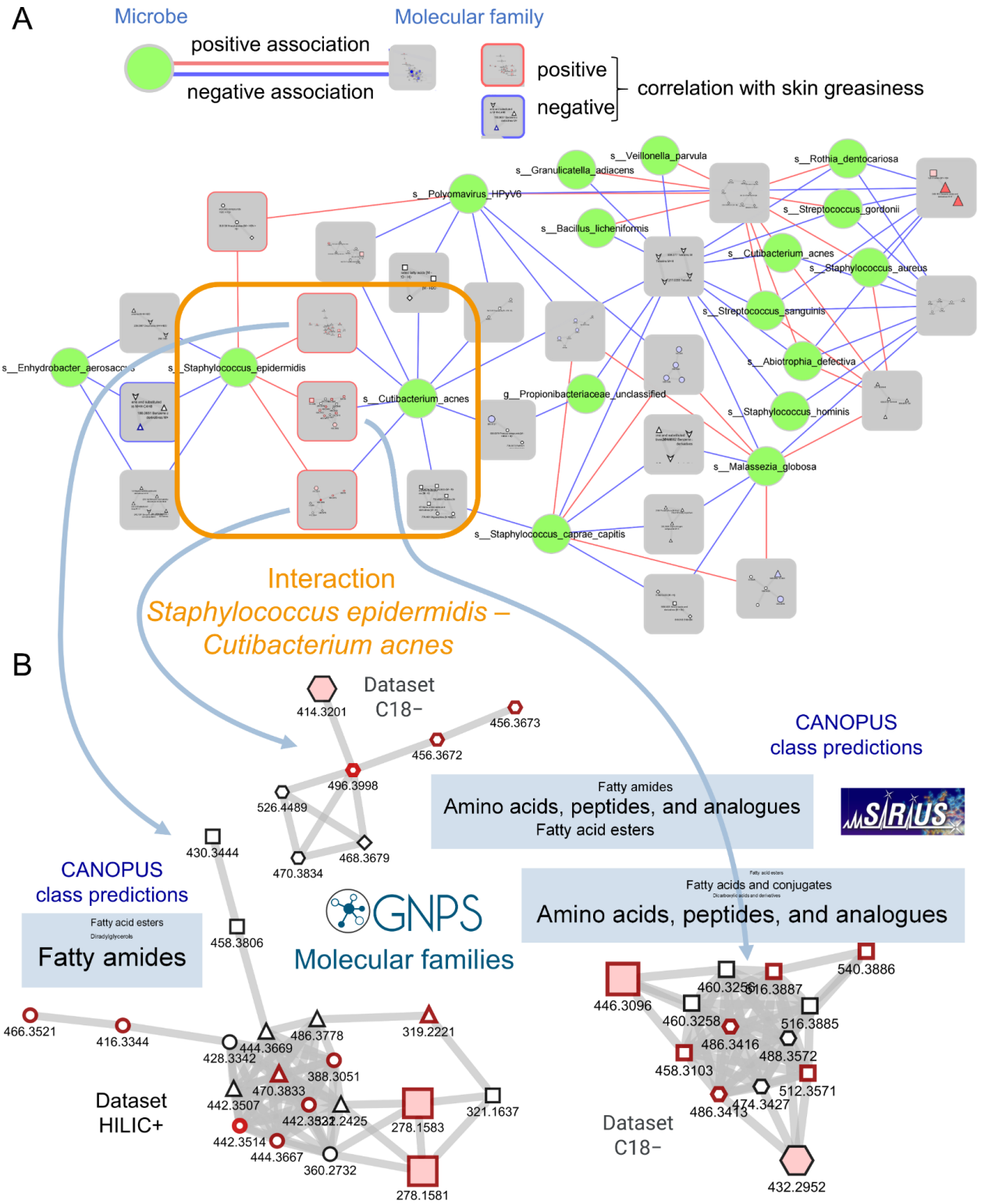


Figure 4. Metabolomics–microbiome data integration: A) Interactions between microbes and GNPS molecular families from MMvec. B) Close-up view on the three molecular families involved in the antagonistic interaction between *S. epidermidis* and *C. acnes*.

METHODS

Recruitment and sample preparation

Recruitment and sampling of study participants were conducted in August 2018 by BASF at the Tarrytown Consumer Testing Center (Tarrytown, NY, U.S.A). The study participants were 19–66 years old, male or female, healthy, and of different ethnic backgrounds. Inclusion criterion for the oily scalp group was a Sebumeter value higher than 120 µg/cm² on the scalp. Inclusion criterion for the normal scalp group was a Sebumeter value between 60 and 120 µg/cm². Samples for metabolomics (oily scalp, n = 41; normal scalp n = 33) and metagenomic sequencing (oily scalp, n = 27; normal scalp n = 29) were taken from the same study participants during the same visit. The lower sample number for metagenomic sequencing is due to some samples failing the quality check after DNA library preparation.

Samples for metagenomic sequencing were collected from the participant's scalp using a sterile cotton swab. Two line segments were created (one on each side of the participant's head) by separating the hair fibers with a comb. A sterile cotton swab was moistened with SCF-1 solution and rubbed along the line segment for 60 seconds. One swab was used for each line segment. Both swabs were then inserted into the same tube containing phosphate buffered saline (PBS) to release the microbes. The tubes were stored on ice until DNA extraction with the QIAamp DNA Microbiome Kit (QIAGEN). The manufacturer's instructions were followed with the exception that the procedure was interrupted after step 3. Samples were stored at −20°C until the procedure was continued with step 4.

Samples for metabolomics were collected from the participant's scalp using HydraFlock swabs (Puritan) following the previously described protocol (10, 28). A line segment was created in the middle of the participant's head by separating the hair fibers with a comb. A swab was moistened with ethanol/water 1:1 (v/v) and rubbed along the line segment for 60 seconds. The tip of the swab was cut and stored in a safe-lock Eppendorf tube at −80°C. The samples were shipped on dry ice to BASF Metabolome Solutions GmbH (Germany), where the metabolomics analysis was performed.

Metagenomics sequencing and analysis.

Library preparation and sequencing. Whole metagenome sequencing libraries were prepared from 26 µL of DNA solution using the NEBNext Ultra II FS DNA Library Prep Kit (New England Biolabs). The DNA was purified and size selected to remove excess adaptors and adaptor dimers using Ampure XP beads (Beckman Coulter Life Sciences). Adaptor-ligated DNA was then subjected to PCR amplification with universal primers and an index barcode that was unique for each sample. The PCR products were cleaned and size-selected with beads to get a 250–600 bp library. The library was checked for quality and quantity on a Bioanalyzer (Agilent) before shipping to a commercial service provider for paired end sequencing on a HiSeq 3000 (Illumina).

Microbiome data processing. After de-multiplexing and quality check, contaminants, adaptors, and known artifacts were removed from the raw reads using bbduk (29). Quality trimming was applied to the ends of reads targeting a Q10 Phred score using bbduk. Reads shorter than 70 bp after trimming were removed. Human reads were removed from the dataset by mapping to the GRCh38 assembly using Bowtie 2 (30) with the --sensitive option. All remaining reads were processed using MetaPhlAn2 (31) to generate a

taxonomic abundance profile for each sample. The resulting table of relative abundances as well as the corresponding sample metadata can be found in the Supplementary Information.

Metabolomics analysis

Sample extraction. The swab tips were first extracted with 500 µL of ethanol, followed by two extractions with 500 µL of 2-propanol. Each extraction comprised the steps of vortexing, shaking for 5 minutes at 1400 rpm, 12°C in a ThermoMixer (Eppendorf), and centrifuging for 5 minutes at 12000 rpm, 12°C. The three extracts from each sample were combined and evaporated in an IR Dancer (Hettlab). The samples were reconstituted with a Bead Ruptor (Omni International) in 1400 µL of methanol/dichloromethane 2:1 (v/v), 120 µL water, 8 µL formic acid, 25 µL polar internal standard solution in water, and 100 µL lipid internal standard solution in methanol/dichloromethane 2:1 (v/v). 100 µL from each sample was transferred into a glass vial for LC-MS analysis. In addition, an equal amount from each sample was taken, pooled, split into the required number of Pool (technical reference) samples, and included with the LC-MS analysis. Blank samples were generated as further technical controls by extracting tips from unused swabs.

LC-MS/MS data acquisition was performed on a TripleTOF 6600 (AB Sciex) by ESI in both positive and negative ionization mode. Each sample was analyzed using both HILIC and RP18 chromatography, giving a total of four runs per sample. The HILIC gradient elution was performed with acetonitrile/water/acetic acid 99:1:0.2 (v/v/v, solvent A) and 0.007 M aqueous ammonium acetate/acetic acid 100:0.2 (v/v, solvent B) from 0% B to 90% B in 5 minutes at a flow of 600 µl/minute. The RP18 gradient elution was performed with water/methanol/0.1 M aqueous ammonium formate/formic acid 49:49:1:0.5 (w/w/w/w, solvent A) and methyl-tert-butylether/2-propanol/methanol/0.1 M aqueous ammonium formate/formic acid 56:28:14:1:0.5 (w/w/w/w, solvent B) from 0% B to 25% B in 0.5 minutes, then to 90% B in 5.4 minutes.

LC-MS/MS data processing. The mass spectrometry data were centroided and converted from the Sciex proprietary format (.wiff) to the open .mzML format using ProteoWizard 3.0 (MSConvert tool) (32). The mzML files were then processed with MZmine toolbox (11) (version 2.37_corr_17.7_kai_merge_2, available on https://github.com/robinschmid/mzmine2/) using the Ion Identity Molecular Networking workflow (33) that allows advanced feature detection for adduct/isotopologue annotations. The MZmine processing was performed on Ubuntu 18.04 LTS 64-bits workstation (Intel Xeon E5-2637, 3.5 GHz, 8 cores, 64 Gb of RAM). The MZmine project, the MZmine batch file (.XML format), and results files (.MGF and .CSV) are available in the MassIVE dataset (MSV000088283). The MZmine batch files contain all the parameters used during the processing for each datasets. In brief, feature detection and deconvolution was performed with the ADAP chromatogram builder (34), and local minimum search algorithm. The isotopologues were regrouped, and the features (specific m/z with retention time peaks) were aligned across samples. The aligned feature list was gap-filled, i.e., a secondary feature detection step that fills in signals from raw data into gaps of missing features. Filters only retained features with an associated fragmentation spectrum and that occurred in a minimum of 2 files. Retention time peak shape correlation grouped features originating from the same molecule and annotated adduct/isotopologue for SIRIUS. Finally the feature quantification table results (.CSV) and fragmentation spectral information (.MGF) were exported with the GNPS module for Feature-Based Molecular Networking (FBMN) analysis on GNPS (13) and with SIRIUS export modules (merge MS/MS across samples with weighted average, in sum

intensity mode, and 10 ppm maximum tolerance, min cosine 0.7, isolation width 5 Da). Other parameters are available in the MZmine batch file.

LC-MS/MS data annotation. The results files from MZmine (.MGF and .CSV files) were uploaded to GNPS (https://gnps.ucsd.edu) (9) and analyzed with the feature-based molecular networking workflow (13). The molecular network topology uses the default parameters (max cluster size = 100, top K = 10, top K filter=on). Spectral library matching was performed against public fragmentation spectra (MS^2) libraries on GNPS and the NIST17 library. Molecular networks were visualized using Cytoscape 3.8.1 software (35).

The MS^2 spectra were systematically annotated with the SIRIUS computational tools (4) computational annotation tool (v. 4.5.3, headless, Linux). Molecular formulas were computed with the SIRIUS module by matching the experimental and predicted isotopic patterns (36), and from fragmentation tree analysis (37) of MS^2.

Molecular formula prediction was refined with the ZODIAC module by using Gibbs sampling (14) for fragmentation spectra that were not chimeric spectra or had poor fragmentation.

In silico structure annotation using structures from bio databases was done with CSI:FingerID (15). Systematic class annotations were obtained with CANOPUS (8) and used the ClassyFire ontology (38). The parameters for SIRIUS tools were set as follows, for SIRIUS: molecular formula candidates retained (100), molecular formula database (ALL), maximum precursor ion *m/z* computed (750), profile (qtof), *m/z* maximum deviation (25 ppm), ions annotated with MZmine were prioritized and other ions were considered in positive ionisation mode ([M+H3N+H]+, [M+H]+, [M+K]+,[M+Na]+, [M+H-H2O]+, [M+H-H4O2]+, [M+NH4]+) and in negative ionisation mode ([M-H]-,[M+Cl]-,[M-H2O-H]-,[M-2H2O-H]-,[M+Br]-); for ZODIAC: threshold filter (0.9), minimum local connections (0); for CSI:FingerID: *m/z* maximum deviation (10 ppm) and biological database (BIO), for CANOPUS: default parameters.

The microbeMASST tool is still in development but is available through a web dashboard (https://masst.gnps2.org/microbemasst; in testing phase) and as a Python script for batch processing. Matching against samples from >60,000 microbial monocultures was performed with a Python script using the .mgf spectra file from the IIMN workflow. Matches were achieved with 0.05 precursor m/z tolerance and fragment m/z tolerance, 0.7 cosine threshold, 4 minimum matched signals, and analog search off. The wider m/z tolerances were chosen to also include datasets of instruments with lower resolving powers, represented within public data. Results were overlaid with the molecular networkings using the feature ID provided by MZmine. Briefly, microbeMASST produced tables of spectral matches to specific data files in the public domain which were then mapped to the curated list of samples with microbe metadata. Those tabular results are then visualized onto trees of the NCBI taxonomy. Interactive HTML/javascript files summarize the results and provide links to other GNPS/MassIVE tools and services, e.g., matching GNPS reference spectra, matching MassIVE datasets, interactive visualization and analysis of MS data files in the GNPS dashboard (39), and spectral mirror plots with the Universal Spectrum Identifier (USI) (40) and the metabolomics spectrum resolver (41). One conclusion from a microbeMASST to a specific NCBI taxonomy is that the target spectrum was previously acquired in another MassIVE dataset.

Statistical analysis of LC-MS/MS data. The feature quantification tables from MZmine were filtered to exclude low-quality features based on the quartiles Q1, Q2, and Q3 of the feature intensity distribution across samples. The quartiles were calculated separately for each feature in the Swab, the Blank, and the Pool samples. A feature was excluded for

high Blank value if the boxes (interquartile ranges) of the Blank and the Pool samples were overlapping in a box plot, i.e., $Q3_{Blank} > Q1_{Pool}$. A feature was excluded for high Pool variability if the box of the Pool samples was wider than 0.5 in a box plot on $\log_{10}$ scale ($IQR_{log10} = \log_{10}(Q3_{Pool}) - \log_{10}(Q1_{Pool}) = \log_{10}(Q3_{Pool} / Q1_{Pool}) > 0.5$).

The intensities from the 44,570 remaining features (all 4 datasets combined) were $\log_{10}$-transformed and normalized to the median of the intensities from the same feature in the Pool samples that were analyzed on the same day. As expected, the resulting intensities showed an approximate normal distribution centered around zero. The intensities were further normalized to the median of the values from all features within each sample and dataset to compensate for differences in the amount of sampled material. The resulting values were used for all downstream statistical analyses, including the MO integration.

The correlation between the $\log_{10}$-transformed and normalized feature intensities and Sebumeter values was calculated using the *stats::lm* function in R (42) with age, gender, ethnic group, and skin moisture (Skicon value) as covariates.

The annotation results from GNPS and SIRIUS were combined to generate a feature annotation metadata file. The statistical analysis and feature annotation results were formatted and mapped into their respective molecular networks for enhanced visualization.

Multi-omics analysis

Microbiome input data. Taxonomic abundance profiles generated with MetaPhlAn2 (see above) were used as input for MMvec. The table of relative abundances can be found in the Supplementary Information together with the corresponding sample metadata.

Metabolomic input data. Filtered, $\log_{10}$-transformed, and normalized feature intensities (see above) were subjected to data reduction as follows: Features present in molecular networks with less than three nodes were discarded to focus on the larger molecular families in the datasets. For each of the 1225 molecular networks with at least three nodes, the median value of the feature intensities was calculated within each sample, resulting in a reduced dataset with one data point for each molecular network in each sample. The dataset was back-transformed to a linear scale and used as input for MMvec.

MMvec. Multi-omics integration was performed using MMvec (6), a neural network-based workflow that learns the co-occurrence probability between a specific molecule against a specific microorganism. MMvec was run via Qiime2 (43) and each computation took less than an hour on a linux computer using 8 cores and GPU with the following parameters: *'--p-arm-the-gpu --p-min-feature-count 4 --p-num-testing-examples 5 --p-summary-interval 1 --p-learning-rate 0.001 --p-latent-dim 3 --p-epochs 100000 --p-batch-size 3000 --p-input-prior 1.8 --p-output-prior 1.1'*. The model convergence was examined with the TensorBoard interface. The metabolomics data were selected based either on their significance in the statistical analysis or by summing feature intensities in a molecular family (molecular network) and by filtering out networks with less than three nodes.

Multi-omics network analysis. The microbes-metabolites (MM) networks generated from MMVec were visualized in Cytoscape (35). The MM networks were generated thanks to a custom python script that extracts the top-n highest co-occurrence conditionality scores for microbe-metabolite interactions using the networkx package (44). When MMvec was performed on molecular families, the MM networks were visualized as 'nested' MM networks where each metabolite node represents an entire molecular family.

References

1. Fan Y, Pedersen O. 2021. Gut microbiota in human metabolic health and disease. Nat Rev Microbiol 19:55–71.

2. Byrd AL, Belkaid Y, Segre JA. 2018. The human skin microbiome. Nat Rev Microbiol 16:143–155.

3. Nakatsuji T, Chen TH, Butcher AM, Trzoss LL, Nam S-J, Shirakawa KT, Zhou W, Oh J, Otto M, Fenical W, Gallo RL. 2018. A commensal strain of Staphylococcus epidermidis protects against skin neoplasia. Sci Adv 4:eaao4502.

4. Dührkop K, Fleischauer M, Ludwig M, Aksenov AA, Melnik AV, Meusel M, Dorrestein PC, Rousu J, Böcker S. 2019. SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. Nat Methods 16:299–302.

5. Aksenov AA, da Silva R, Knight R, Lopes NP, Dorrestein PC. 2017. Global chemical analysis of biology by mass spectrometry. Nature Reviews Chemistry 1:1–20.

6. Morton JT, Aksenov AA, Nothias LF, Foulds JR, Quinn RA, Badri MH, Swenson TL, Van Goethem MW, Northen TR, Vazquez-Baeza Y, Wang M, Bokulich NA, Watters A, Song SJ, Bonneau R, Dorrestein PC, Knight R. 2019. Learning representations of microbe-metabolite interactions. Nat Methods 16:1306–1314.

7. Cai Y, Zhou Z, Zhu Z-J. 2023. Advanced analytical and informatic strategies for metabolite annotation in untargeted metabolomics. Trends Analyt Chem 158:116903.

8. Dührkop K, Nothias L-F, Fleischauer M, Reher R, Ludwig M, Hoffmann MA, Petras D, Gerwick WH, Rousu J, Dorrestein PC, Böcker S. 2021. Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. Nat Biotechnol 39, 462–471.

9. Wang M, Carver JJ, Phelan VV, Sanchez LM, Garg N, Peng Y, Nguyen DD, Watrous J, Kapono CA, Luzzatto-Knaan T, Porto C, Bouslimani A, Melnik AV, Meehan MJ, Liu W-T, Crüsemann M, Boudreau PD, Esquenazi E, Sandoval-Calderón M, Kersten RD, Pace LA, Quinn RA, Duncan KR, Hsu C-C, Floros DJ, Gavilan RG, Kleigrewe K, Northen T, Dutton RJ, Parrot D, Carlson EE, Aigle B, Michelsen CF, Jelsbak L, Sohlenkamp C, Pevzner P, Edlund A, McLean J, Piel J, Murphy BT, Gerwick L, Liaw C-C, Yang Y-L, Humpf H-U, Maansson M, Keyzers RA, Sims AC, Johnson AR, Sidebottom AM, Sedio BE, Klitgaard A, Larson CB, P CAB, Torres-Mendoza D, Gonzalez DJ, Silva DB, Marques LM, Demarque DP, Pociute E, O'Neill EC, Briand E, Helfrich EJN, Granatosky EA, Glukhov E, Ryffel F, Houson H, Mohimani H, Kharbush JJ, Zeng Y, Vorholt JA, Kurita KL, Charusanti P, McPhail KL, Nielsen KF, Vuong L, Elfeki M, Traxler MF, Engene N, Koyama N, Vining OB, Baric R, Silva RR, Mascuch SJ, Tomasi S, Jenkins S, Macherla V, Hoffman T, Agarwal V, Williams PG, Dai J, Neupane R, Gurr J, Rodríguez AMC, Lamsa A, Zhang C, Dorrestein K, Duggan BM, Almaliti J, Allard P-M, Phapale P, Nothias L-F, Alexandrov T, Litaudon M, Wolfender J-L, Kyle JE, Metz TO, Peryea T, Nguyen D-T, VanLeer D, Shinn P, Jadhav A, Müller R, Waters KM, Shi W, Liu X, Zhang L, Knight R, Jensen PR, Palsson BO, Pogliano K, Linington RG, Gutiérrez M, Lopes NP,

Gerwick WH, Moore BS, Dorrestein PC, Bandeira N. 2016. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. Nat Biotechnol 34:828–837.

10. Bouslimani A, Porto C, Rath CM, Wang M, Guo Y, Gonzalez A, Berg-Lyon D, Ackermann G, Christensen GJM, Nakatsuji T, Zhang L, Borkowski AW, Meehan MJ, Dorrestein K, Gallo RL, Bandeira N, Knight R, Alexandrov T, Dorrestein PC. 2015. Molecular cartography of the human skin surface in 3D. Proc Natl Acad Sci U S A 112:2120–2129.

11. Pluskal T, Castillo S, Villar-Briones A, Oresic M. 2010. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC Bioinformatics 11:395.

12. Schmid R, Petras D, Nothias L-F, Wang M, Aron AT, Jagels A, Tsugawa H, Rainer J, Garcia-Aloy M, Dührkop K, Korf A, Pluskal T, Kameník Z, Jarmusch AK, Caraballo-Rodríguez AM, Weldon KC, Nothias-Esposito M, Aksenov AA, Bauermeister A, Albarracin Orio A, Grundmann CO, Vargas F, Koester I, Gauglitz JM, Gentry EC, Hövelmann Y, Kalinina SA, Pendergraft MA, Panitchpakdi M, Tehan R, Le Gouellec A, Aleti G, Mannochio Russo H, Arndt B, Hübner F, Hayen H, Zhi H, Raffatellu M, Prather KA, Aluwihare LI, Böcker S, McPhail KL, Humpf H-U, Karst U, Dorrestein PC. 2021. Ion identity molecular networking for mass spectrometry-based metabolomics in the GNPS environment. Nat Commun 12:3832.

13. Nothias L-F, Petras D, Schmid R, Dührkop K, Rainer J, Sarvepalli A, Protsyuk I, Ernst M, Tsugawa H, Fleischauer M, Aicheler F, Aksenov AA, Alka O, Allard P-M, Barsch A, Cachet X, Caraballo-Rodriguez AM, Da Silva RR, Dang T, Garg N, Gauglitz JM, Gurevich A, Isaac G, Jarmusch AK, Kameník Z, Kang KB, Kessler N, Koester I, Korf A, Le Gouellec A, Ludwig M, Martin H C, McCall L-I, McSayles J, Meyer SW, Mohimani H, Morsy M, Moyne O, Neumann S, Neuweger H, Nguyen NH, Nothias-Esposito M, Paolini J, Phelan VV, Pluskal T, Quinn RA, Rogers S, Shrestha B, Tripathi A, van der Hooft JJJ, Vargas F, Weldon KC, Witting M, Yang H, Zhang Z, Zubeil F, Kohlbacher O, Böcker S, Alexandrov T, Bandeira N, Wang M, Dorrestein PC. 2020. Feature-based molecular networking in the GNPS analysis environment. Nat Methods 17:905–908.

14. Ludwig M, Nothias L-F, Dührkop K, Koester I, Fleischauer M, Hoffmann MA, Petras D, Vargas F, Morsy M, Aluwihare L, Dorrestein PC, Böcker S. 2020. Database-independent molecular formula annotation using Gibbs sampling through ZODIAC. Nature Machine Intelligence 2:629–641.

15. Dührkop K, Shen H, Meusel M, Rousu J, Böcker S. 2015. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. Proc Natl Acad Sci U S A 112:12580–12585.

16. Johansson I, Svensson M. 2001. Surfactants based on fatty acids and other natural hydrophobes. Curr Opin Colloid Interface Sci 6:178–188.

17. Kruis AJ, Bohnenkamp AC, Patinios C, van Nuland YM, Levisson M, Mars AE, van den Berg C, Kengen SWM, Weusthuis RA. 2019. Microbial production of short and medium chain esters: Enzymes, pathways, and applications. Biotechnol Adv 37:107407.

18. Ludovici M, Kozul N, Materazzi S, Risoluti R, Picardo M, Camera E. 2018. Influence of the sebaceous gland density on the stratum corneum lipidome. Sci Rep 8:11500.

19. Wang M, Jarmusch AK, Vargas F, Aksenov AA, Gauglitz JM, Weldon K, Petras D, da Silva R, Quinn R, Melnik AV, van der Hooft JJJ, Caraballo-Rodríguez AM, Nothias LF, Aceves CM, Panitchpakdi M, Brown E, Di Ottavio F, Sikora N, Elijah EO, Labarta-Bajo L, Gentry EC, Shalapour S, Kyle KE, Puckett SP, Watrous JD, Carpenter CS, Bouslimani A, Ernst M, Swafford AD, Zúñiga EI, Balunas MJ, Klassen JL, Loomba R, Knight R, Bandeira N, Dorrestein PC. 2020. Mass spectrometry searches using MASST. Nat Biotechnol 38:23–26.

20. Christensen GJM, Brüggemann H. 2014. Bacterial skin commensals and their role as host guardians. Benef Microbes 5:201–215.

21. Christensen GJM, Scholz CFP, Enghild J, Rohde H, Kilian M, Thürmer A, Brzuszkiewicz E, Lomholt HB, Brüggemann H. 2016. Antagonism between Staphylococcus epidermidis and Propionibacterium acnes and its genomic basis. BMC Genomics 17:152.

22. Fournière M, Latire T, Souak D, Feuilloley MGJ, Bedoux G. 2020. Staphylococcus epidermidis and Cutibacterium acnes: Two Major Sentinels of Skin Microbiota and the Influence of Cosmetics. Microorganisms 8.

23. Rosenstein R, Götz F. 2000. Staphylococcal lipases: biochemical and molecular characterization. Biochimie 82:1005–1014.

24. Liu L-P, Huang L-H, Ding X-T, Yan L, Jia S-R, Dai Y-J, Xie Y-Y, Zhong C. 2019. Identification of Quorum-Sensing Molecules of N-Acyl-Homoserine Lactone in Gluconacetobacter Strains by Liquid Chromatography-Tandem Mass Spectrometry. Molecules 24.

25. Turkina MV, Vikström E. 2019. Bacteria-Host Crosstalk: Sensing of the Quorum in the Context of Pseudomonas aeruginosa Infections. J Innate Immun 11:263–279.

26. Parsons JB, Rock CO. 2013. Bacterial lipids: metabolism and membrane homeostasis. Prog Lipid Res 52:249–276.

27. Pappas A. 2009. Epidermal surface lipids. Dermatoendocrinol 1:72–76.

28. Protsyuk I, Melnik AV, Nothias L-F, Rappez L, Phapale P, Aksenov AA, Bouslimani A, Ryazanov S, Dorrestein PC, Alexandrov T. 2018. 3D molecular cartography using LC-MS facilitated by Optimus and ’ili software. Nat Protoc 13:134–154.

29. Bushnell B. 2014. BBMap: a fast, accurate, splice-aware aligner. Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States).

30. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359.

31. Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, Tett A, Huttenhower C, Segata N. 2015. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat Methods 12:902–903.

32. Chambers MC, Maclean B, Burke R, Amodei D, Ruderman DL, Neumann S, Gatto L, Fischer B, Pratt B, Egertson J, Hoff K, Kessner D, Tasman N, Shulman N, Frewen B, Baker TA, Brusniak M-Y, Paulse C, Creasy D, Flashner L, Kani K, Moulding C, Seymour SL, Nuwaysir LM, Lefebvre B, Kuhlmann F, Roark J, Rainer P, Detlev S, Hemenway T, Huhmer A, Langridge J, Connolly B, Chadick T, Holly K, Eckels J, Deutsch EW, Moritz RL, Katz JE, Agus DB, MacCoss M, Tabb DL, Mallick P. 2012. A cross-platform toolkit for mass spectrometry and proteomics. Nat Biotechnol 30:918–920.

33. Schmid R, Petras D, Nothias L-F, Wang M, Aron AT, Jagels A, Tsugawa H, Rainer J, Garcia-Aloy M, Dührkop K, Korf A, Pluskal T, Kameník Z, Jarmusch AK, Caraballo-Rodríguez AM, Weldon K, Nothias-Esposito M, Aksenov AA, Bauermeister A, Orio AA, Grundmann CO, Vargas F, Koester I, Gauglitz JM, Gentry EC, Hövelmann Y, Kalinina SA, Pendergraft MA, Panitchpakdi MW, Tehan R, Le Gouellec A, Aleti G, Russo HM, Arndt B, Hübner F, Hayen H, Zhi H, Raffatellu M, Prather KA, Aluwihare LI, Böcker S, McPhail KL, Humpf H-U, Karst U, Dorrestein PC. 2020. Ion Identity Molecular Networking in the GNPS Environment. Cold Spring Harbor Laboratory.

34. Du X, Smirnov A, Pluskal T, Jia W, Sumner S. 2020. Metabolomics Data Preprocessing Using ADAP and MZmine 2. Methods Mol Biol 2104:25–48.

35. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13:2498–2504.

36. Böcker S, Letzel MC, Lipták Z, Pervukhin A. 2009. SIRIUS: decomposing isotope patterns for metabolite identification. Bioinformatics 25:218–224.

37. Böcker S, Dührkop K. 2016. Fragmentation trees reloaded. J Cheminform 8:5.

38. Kim H, Wang M, Leber C, Nothias L-F, Reher R, Kang KB, van der Hooft JJJ, Dorrestein P, Gerwick W, Cottrell G. 2020. NPClassifier: A Deep Neural Network-Based Structural Classification Tool for Natural Products. ChemRxiv.

39. Petras D, Phelan VV, Acharya D, Allen AE, Aron AT, Bandeira N, Bowen BP, Belle-Oudry D, Boecker S, Cummings DA Jr, Deutsch JM, Fahy E, Garg N, Gregor R, Handelsman J, Navarro-Hoyos M, Jarmusch AK, Jarmusch SA, Louie K, Maloney KN, Marty MT, Meijler MM, Mizrahi I, Neve RL, Northen TR, Molina-Santiago C, Panitchpakdi M, Pullman B, Puri AW, Schmid R, Subramaniam S, Thukral M, Vasquez-Castro F, Dorrestein PC, Wang M. 2022. GNPS Dashboard: collaborative exploration of mass spectrometry data in the web browser. Nat Methods 19:134–136.

40. Deutsch EW, Perez-Riverol Y, Carver J, Kawano S, Mendoza L, Van Den Bossche T, Gabriels R, Binz P-A, Pullman B, Sun Z, Shofstahl J, Bittremieux W, Mak TD, Klein J, Zhu Y, Lam H, Vizcaíno JA, Bandeira N. 2021. Universal Spectrum Identifier for mass spectra. Nat Methods 18:768–770.

41. Bittremieux W, Chen C, Dorrestein PC, Schymanski EL, Schulze T, Neumann S, Meier R, Rogers S, Wang M. 2020. Universal MS/MS Visualization and Retrieval with the Metabolomics Spectrum Resolver Web Service. bioRxiv.

42. Ripley BD. 2001. The R project in statistical computing. MSOR Connections The newsletter of the LTSN Maths.

43. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ, Brown CT, Callahan BJ, Caraballo-Rodríguez AM, Chase J, Cope EK, Da Silva R, Diener C, Dorrestein PC, Douglas GM, Durall DM, Duvallet C, Edwardson CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibbons SM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S, Holste H, Huttenhower C, Huttley GA, Janssen S, Jarmusch AK, Jiang L, Kaehler BD, Kang KB, Keefe CR, Keim P, Kelley ST, Knights D, Koester I, Kosciolek T, Kreps J, Langille MGI, Lee J, Ley R, Liu Y-X, Loftfield E, Lozupone C, Maher M, Marotz C, Martin BD, McDonald D, McIver LJ, Melnik AV, Metcalf JL, Morgan SC, Morton JT, Naimey AT, Navas-Molina JA, Nothias LF, Orchanian SB, Pearson T, Peoples SL, Petras D, Preuss ML, Pruesse E, Rasmussen LB, Rivers A, Robeson MS 2nd, Rosenthal P, Segata N, Shaffer M, Shiffer A, Sinha R, Song SJ, Spear JR, Swafford AD, Thompson LR, Torres PJ, Trinh P, Tripathi A, Turnbaugh PJ, Ul-Hasan S, van der Hooft JJJ, Vargas F, Vázquez-Baeza Y, Vogtmann E, von Hippel M, Walters W, Wan Y, Wang M, Warren J, Weber KC, Williamson CHD, Willis AD, Xu ZZ, Zaneveld JR, Zhang Y, Zhu Q, Knight R, Caporaso JG. 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol 37:852–857.

44. Hagberg A, Swart P, S Chult D. 2008. Exploring network structure, dynamics, and function using NetworkX. Los Alamos National Lab.(LANL), Los Alamos, NM (United States).

Author contributions statement

L.-F.N., P.C.D., R.F., and P.T. conceptualized the multi-omics strategy. A.G., S. L.-O., and V.A.. conceptualized the scalp study, L.-F.N. performed the mass spectrometry data processing and mass spectra annotation. P.T. performed the statistical analysis. L.-F.N. and P.T. interpreted the mass spectrometry data and performed the multi-omics integration using MMvec. R.S. performed the microbeMASST analysis. H.C. performed metagenomics sequencing and annotation. L.-F.N. and P.T. wrote the manuscript. P.C.D. and P.T. supervised the work presented in this manuscript.

Competing interests statement

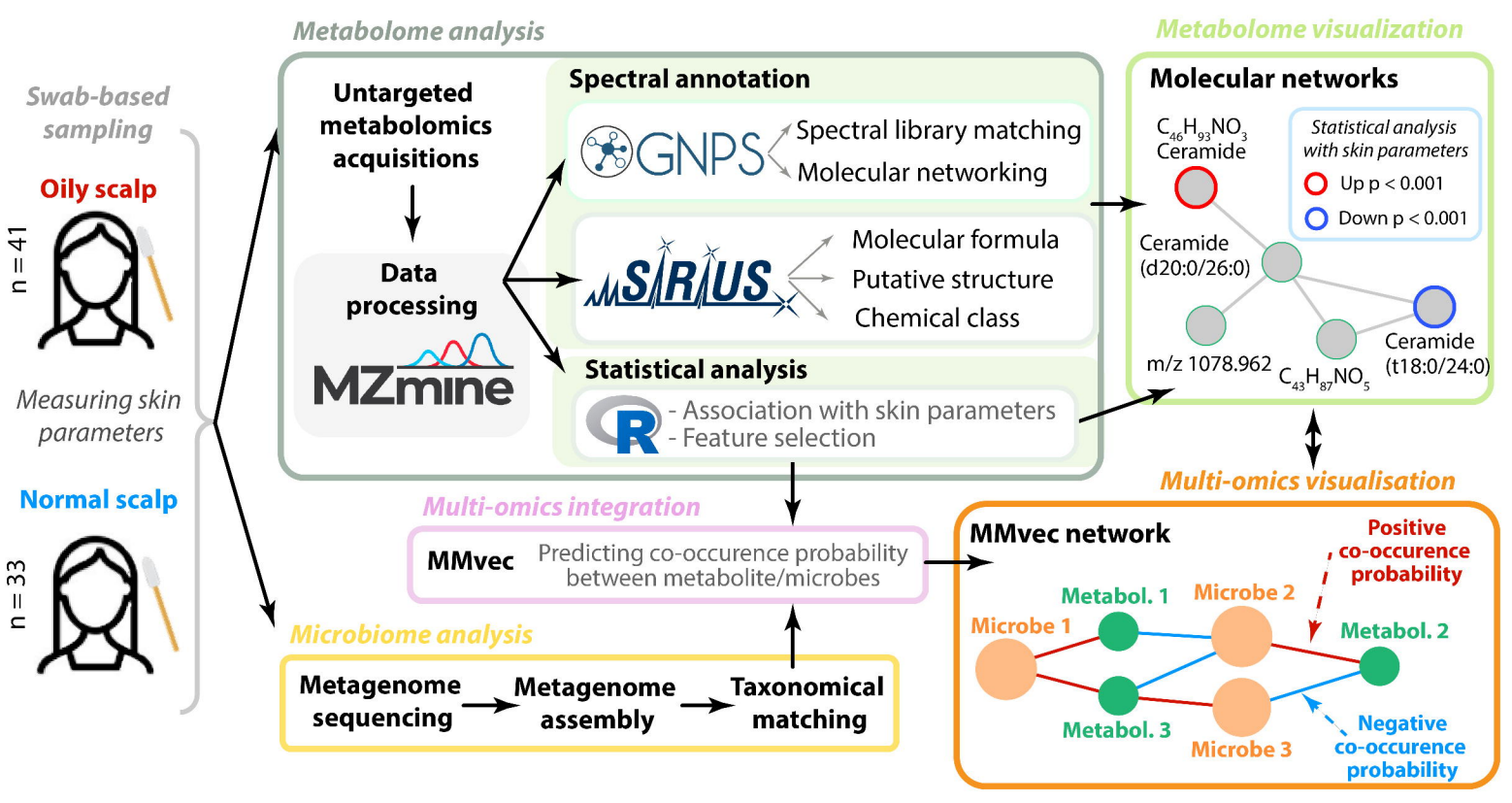
PCD is an advisor to Cybele and co-founder and scientific advisor to Ometa and Enveda, with prior approval by UC San Diego.

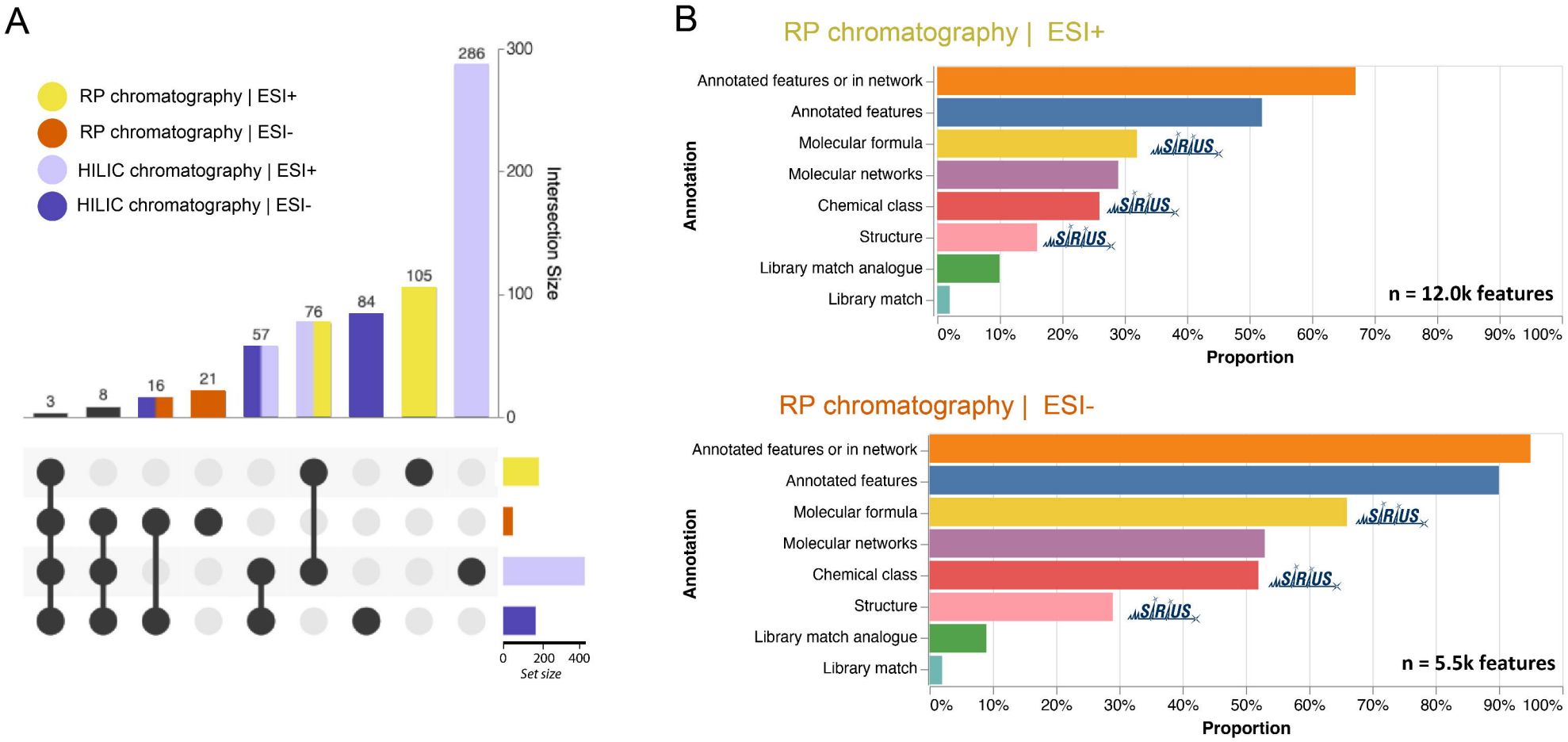
Ethics statement

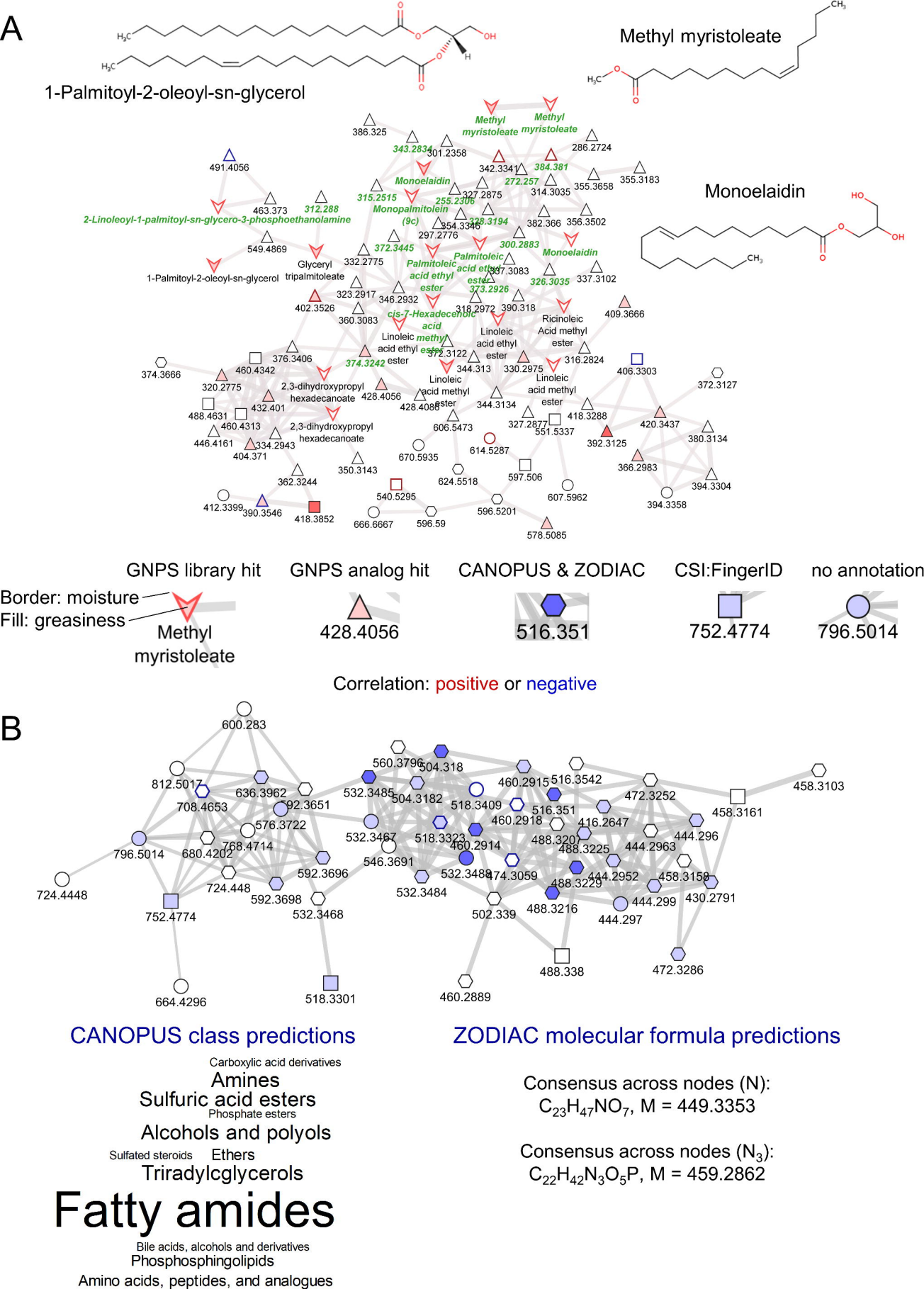
The protocol for the present study (# TC-0118-003-095 A and B) is based on Parent Protocol "Hair/Scalp Study" (# TC-0315-001-006), which was reviewed and approved by the BASF Tarrytown Consumer Testing Center Institutional Review Board (TCTC IRB). The protocol also covers the collection, sequencing, and data analysis of microbiome samples. Informed consent was obtained from all participating volunteers and the experiments conformed to the principles set out in the WMA Declaration of Helsinki and the Department of Health and Human Services Belmont Report.

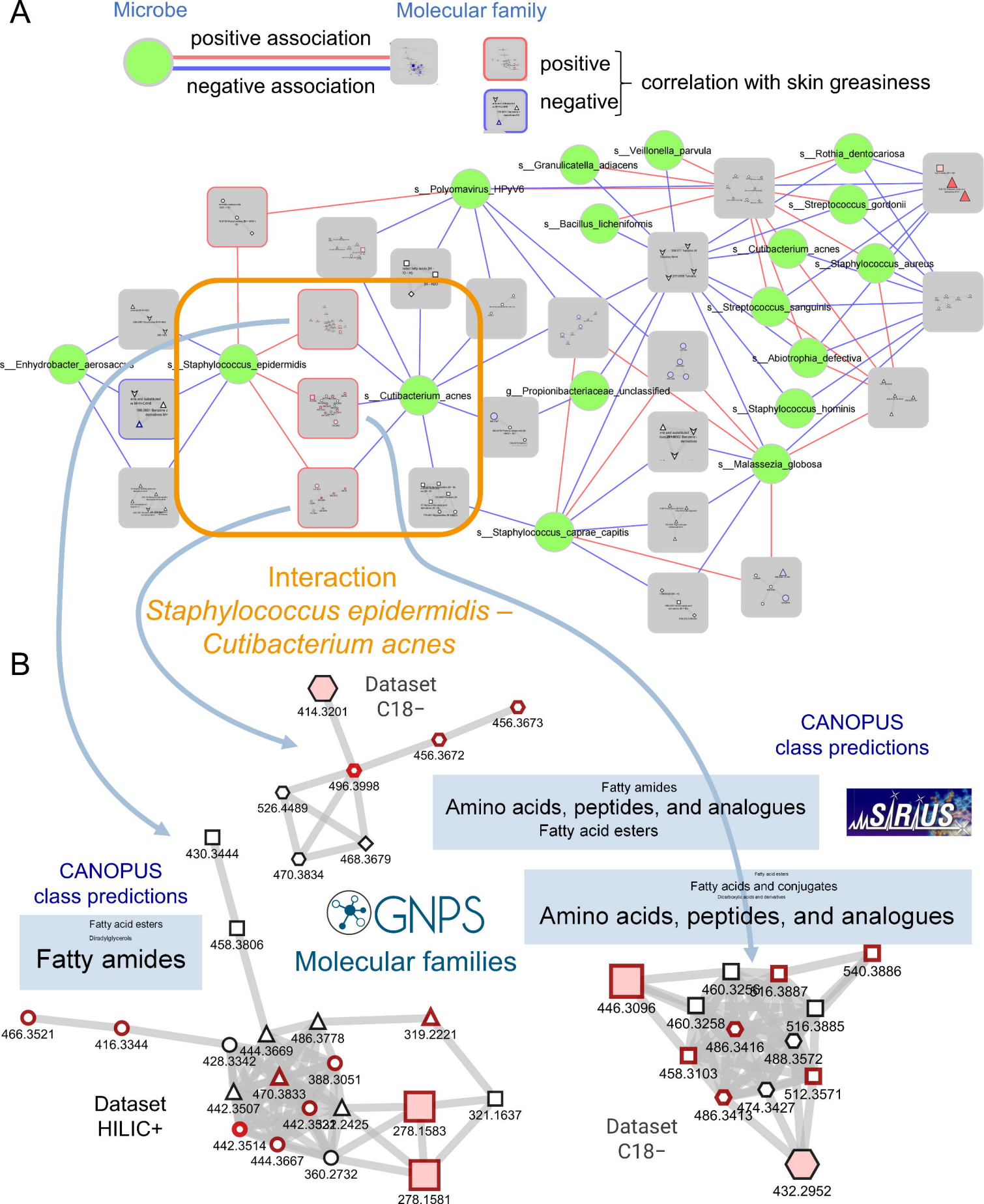
Acknowledgements

The authors thank Ina Erbe, Nicole Rocker, and Peter Driemert at BASF Metabolome Solutions GmbH for the generation of the mass spectrometry data, James T. Morton at the University of California, San Diego, for help with the MMvec analysis, and the BASF California Research Alliance (CARA) for funding.


Metabolome analysis
Metabolome visualization
Swab-based sampling
Oily scalp
n = 41
Measuring skin parameters
Normal scalp
n = 33
Untargeted metabolomics acquisitions
Data processing
MZmine
Spectral annotation
GNPS
Spectral library matching
Molecular networking
SIRIUS
Molecular formula
Putative structure
Chemical class
Statistical analysis
R
- Association with skin parameters
- Feature selection
Molecular networks
$C_{46}H_{93}NO_3$
Ceramide
Statistical analysis with skin parameters
Up p < 0.001
Down p < 0.001
Ceramide (d20:0/26:0)
m/z 1078.962
$C_{43}H_{87}NO_5$
Ceramide (t18:0/24:0)
Multi-omics integration
MMvec
Predicting co-occurence probability between metabolite/microbes
Multi-omics visualisation
MMvec network
Positive co-occurence probability
Metabol. 1
Microbe 2
Microbe 1
Metabol. 2
Metabol. 3
Microbe 3
Negative co-occurence probability
Microbiome analysis
Metagenome sequencing
Metagenome assembly
Taxonomical matching


A
RP chromatography | ESI+
RP chromatography | ESI-
HILIC chromatography | ESI+
HILIC chromatography | ESI-
286
105
84
76
57
21
16
8
3
300
200
100
0
Intersection Size
0
200
400
Set size
B
RP chromatography | ESI+
Annotation
Annotated features or in network
Annotated features
Molecular formula
Molecular networks
Chemical class
Structure
Library match analogue
Library match
n = 12.0k features
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Proportion
RP chromatography | ESI-
Annotation
Annotated features or in network
Annotated features
Molecular formula
Molecular networks
Chemical class
Structure
Library match analogue
Library match
n = 5.5k features
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Proportion


A
1-Palmitoyl-2-oleoyl-sn-glycerol
Methyl myristoleate
Monoelaidin
GNPS library hit
GNPS analog hit
CANOPUS & ZODIAC
CSI:FingerID
no annotation
Border: moisture
Fill: greasiness
Methyl myristoleate
428.4056
516.351
752.4774
796.5014
Correlation: positive or negative
B
CANOPUS class predictions
Carboxylic acid derivatives
Amines
Sulfuric acid esters
Phosphate esters
Alcohols and polyols
Sulfated steroids
Ethers
Triradylcglycerols
Fatty amides
Bile acids, alcohols and derivatives
Phosphosphingolipids
Amino acids, peptides, and analogues
ZODIAC molecular formula predictions
Consensus across nodes (N):
$C_{23}H_{47}NO_7$, M = 449.3353
Consensus across nodes (N_3):
$C_{22}H_{42}N_3O_5P$, M = 459.2862


A
Microbe
Molecular family
positive association
negative association
positive
negative
correlation with skin greasiness
s__Veillonella_parvula
s__Granulicatella_adiacens
s__Rothia_dentocariosa
s__Polyomavirus_HPyV6
s__Streptococcus_gordonii
s__Bacillus_licheniformis
s__Cutibacterium_acnes
s__Staphylococcus_aureus
s__Streptococcus_sanguinis
s__Enhydrobacter_aerosaccus
s__Staphylococcus_epidermidis
s__Abiotrophia_defectiva
s__Cutibacterium_acnes
g__Propionibacteriaceae_unclassified
s__Staphylococcus_hominis
s__Malassezia_globosa
s__Staphylococcus_caprae_capitis
Interaction
Staphylococcus epidermidis –
Cutibacterium acnes
B
Dataset
C18−
414.3201
456.3673
456.3672
496.3998
526.4489
468.3679
470.3834
CANOPUS
class predictions
Fatty amides
Amino acids, peptides, and analogues
Fatty acid esters
SIRIUS
430.3444
458.3806
CANOPUS
class predictions
Fatty acid esters
Diradylglycerols
Fatty amides
GNPS
Molecular families
Fatty acid esters
Fatty acids and conjugates
Dicarboxylic acids and derivatives
Amino acids, peptides, and analogues
466.3521
416.3344
444.3669
428.3342
486.3778
319.2221
388.3051
470.3833
442.3507
442.3514
444.3667
360.2732
278.1583
321.1637
278.1581
Dataset
HILIC+
446.3096
460.3256
516.3887
540.3886
460.3258
516.3885
486.3416
488.3572
458.3103
512.3571
474.3427
486.3413
432.2952
Dataset
C18−