An enhanced characterization of the human skin microbiome: a new biodiversity of microbial interactions

Akintunde Emiola¹, Wei Zhou¹, Julia Oh^{1*}

¹The Jackson Laboratory for Genomic Medicine, Farmington, Connecticut, USA

*Corresponding author. Julia.Oh@jax.org

ABSTRACT

The healthy human skin microbiome is shaped by skin site physiology, individual-specific factors, and is largely stable over time despite significant environmental perturbation. Studies identifying these characteristics used shotgun metagenomic sequencing for high resolution reconstruction of the bacteria, fungi, and viruses in the community. However, these conclusions were drawn from a relatively small proportion of the total sequence reads analyzable by mapping to known reference genomes. ‘Reference-free’ approaches, based on *de novo* assembly of reads into genome fragments, are also limited in their ability to capture low abundance species, small genomes, and to discriminate between more similar genomes. To account for the large fraction of non-human unmapped reads on the skin—referred to as microbial ‘dark matter’—we used a hybrid *de novo* and reference-based approach to annotate a metagenomic dataset of 698 healthy human skin samples. This approach reduced the overall proportion of uncharacterized reads from 42% to 17%. With our refined characterization, we revisited assumptions about the skin microbiome, and demonstrated higher biodiversity and lower stability, particularly in dry and moist skin sites. To investigate hypotheses underlying stability, we examined growth dynamics and interspecies interactions in these communities. Surprisingly, even though most skin sites were relatively stable, many dominant skin microbes, including *Cutibacterium acnes* and staphylococci, were actively growing in the skin, with poor or no relationship between growth rate and relative abundance, suggesting that host selection or interspecies competition may be important factors maintaining community homeostasis. To investigate other mechanisms facilitating adaptation to a specific skin site, we identified *Staphylococcus epidermidis* genes that are likely involved in stress response and provide mechanisms essential for growth in oily sites. Finally, horizontal gene transfer—another mechanism of competition by which strains may swap antagonistic or virulent coding regions—was relatively limited in healthy skin, but suggested exchange of different metabolic and environmental tolerance pathways. Altogether, our findings underscore the value of a combined reference-based and *de novo* approach to provide significant new insights into microbial composition, physiology, and interspecies interactions to maintain community homeostasis in the healthy human skin microbiome.

BACKGROUND

Deep metagenomic shotgun sequencing is a powerful tool to interrogate composition and function of complex microbial communities. Microbial communities offer the potential for discovery of a tremendous suite of previously unknown biological functions, for example, new bioactive

compounds, antimicrobials, virulence factors, or metabolic pathways. Such discovery has relied on the ability to survey and deconvolute species from mixed microbial consortia. Advances in next-generation sequencing and computational analyses have, in recent years, greatly furthered efforts to reconstruct microbial communities at the species^{1,2}, strain^{2,3}, and even single nucleotide polymorphism level^{3,4}, examining function, transmission, and stability of the resident microbes.

However, interpretations of many metagenomic datasets are limited by the inability to characterize a large fraction of the total microbial reads present in the original sample^{5,6}. This uncharacterized sequence space, or microbial ‘dark matter’⁷, typically results from the inability to map a sequence read to a known microbial reference genome and can exceed 96% of sequence reads within a sample⁵. Such ‘reference-based’ approaches, whether mapping reads to complete genomes⁸ or marker genes⁹, have high sensitivity and discriminatory ability between even very similar genomes⁸. However, microbes with no representative reference, or those with significant pangenomic variation, which can account for considerable within-species diversity in gene content¹⁰, are not captured. Conversely, reference-free approaches based on *de novo* assembly to aggregate reads into longer stretches of contiguous DNA sequence, can aid in the identification and characterization of new genomes. However, *de novo* assembly-based approaches are less effective in capturing small genomes (e.g., viruses), low-abundance microbes, and in discriminating between very similar genomes.

By combining both approaches into a holistic framework, we aimed to reduce the proportion of uncharacterized sequence space in a metagenomics dataset, and thus provide new insights into the biological function and interspecies interactions of these microbial communities. We used a hybrid *de novo* and reference-based approach aimed at characterizing microbial dark matter in the skin metagenome. Our previous analyses of this dataset (698 samples), which were exclusively reference-based, showed that the skin microbiome is defined primarily by the physiological characteristics of the skin site (e.g., whether it was a sebaceous, moist, dry, or foot site), then by host-intrinsic factors that confer individuality in strain representation and the presence of low-abundance and transient organisms^{5,6}. More intriguing was the observation that the skin microbiome is remarkably stable even over years, despite the exposure of skin to different hygiene practices and the external environment⁶. However, our conclusions were based on an incomplete portrait with, on average, half of each sample remaining uncharacterized by our reference-based analyses⁵. By incorporating additional information from microbial dark matter, we stood to gain significant new insights into the landscape of skin biodiversity and microbial stability.

Leveraging our integrated approach, we uncovered previously unaccounted-for biodiversity and reduced microbial stability in the skin microbiome. We used this refined characterization to more deeply probe interspecies interactions, identifying intra-genus diversity and mechanisms underlying stability and inter-species interactions in the skin, including new assessments on growth rate and horizontal gene transfer. Our results demonstrate the highest resolution analysis of the skin microbiome to date, and provide new hypotheses for how skin microbes interact and compete to maintain homeostatic community conditions.

RESULTS

A hybrid *de novo* and reference-based microbial community analysis

To address the significant uncharacterized sequence space (mean $\pm$ sd 42% $\pm$ 24%) in our initial analysis of a 698-sample longitudinal skin metagenomic dataset (Supplementary Fig. 1), we used reference-independent approaches to reconstruct composition. With the improvement of *de novo* assembly algorithms to input large datasets¹¹, we concatenated our samples and assembled iteratively, resulting in 75% $\pm$ 19% reads incorporated into the assemblies (Supplementary Fig. 2). The 1,037,465 resultant contigs >1kb were then grouped into genome 'bins' based on co-abundance clustering and nucleotide composition. However, because this approach is limited in its ability to recover small genomes, low-abundance species, and to ascertain precise taxonomic classifications, we speculated that integrating reference-based analyses (Fig. 1A) would further reduce dark matter beyond the 33% $\pm$ 21% reduction observed by mapping reads to our *de novo* reference set only (Fig. 1B). Microbial reads unmappable to our *de novo* reference catalogue were annotated by mapping to a reference database of fungal, bacterial, and viral genomes. Reference-based and *de novo* classifications were integrated with a normalization step that took into consideration the total proportion of reads derived from each approach (Supplementary Fig. 2). While using *de novo* references significantly aided reconstruction of microbiota, our hybrid approach most considerably reduced the proportion of dark matter (16% $\pm$ 17%; Fig. 1B and C).

A new biodiversity of the human skin metagenome

Our new compositional analysis was largely concordant with our previous findings that the skin microbiome is predominated by *Staphylococcus*, *Cutibacterium* (formerly *Propionibacterium*), *Corynebacterium*, and *Malassezia* species^{5,6} (Fig. 1C, Fig. 1D). However, we uncovered considerably more diversity of *Staphylococcus*, *Corynebacterium*, *Proteobacteria*, and fungal genomes at most skin sites (Fig. 1C, Supplementary Fig. 3A). For example, we identified a previously uncharacterized, but abundant, *Lactobacillales* colonizing the external auditory canal (Ea) as *Alloiococcus otitis*. 9% $\pm$ 13% of reads mapped to bins were unclassifiable based on BLASTn alignment to the NCBI nt database. These likely represented contigs that belong either to uncharacterized genomes or undiscovered pan-genomic variation of a lower abundance strain that could not be binned with its species unit. These otherwise unclassifiable bins were most frequently bacterial, underscoring that the majority of undiscovered biodiversity in the skin is not fungal or viral (Supplementary Fig. 3B).

Our revised classification showed lower representation of *C. acnes* than previous analyses, and markedly lower *Propionibacterium* phage representation in sebaceous regions. *De novo*-only approach did not capture viral contigs; these were most accurately recovered with the combined use of reference genomes, for example in the alar crease (Al) (Fig. 1C).

With our integrated classification, we re-examined our conclusions of diversity and stability at the community level. All skin sites except the ear and foot had higher diversity than previously reported, which is expected given the identification and inclusion of more genomes/genome bins than in our original analyses (Fig. 1E). Because resolved dark matter only represented a few genomes in the external auditory canal (Ea) (mostly *Lactobacillales*) and foot sites (mostly *Staphylococcus* and *Corynebacterium*), diversity was unchanged in these regions. Since previous

analyses correlated increased diversity with decreased stability⁶, we re-evaluated our conclusions on stability over the ~month (T2-T3) and ~year intervals (T1-T2) collected in this study. Most sites were less stable than originally defined (Fig. 1F)—for example, the hand and forearm (dry) sites, which are highly exposed to the external environment (Fig. 1F, Supplementary Fig. 4A). This may be due to behavioral patterns like hand-washing or increased acquisition of transient, environmental microbes. Longitudinal tracking of individual species' dynamics over time showed that community stability is driven by specific microbes (Supplementary Fig. 4B). For example, newly identified *Corynebacterium* species (e.g., *jeikeium*) were associated with stability whereas staphylococci showed more fluctuation over time.

Mechanisms underlying interspecies interactions: growth dynamics of skin microbiota

A significant limitation of previous skin microbiome studies is the absence of information on microbial activity, which is necessary to understand potential underlying homeostatic mechanisms as microbes are unlikely to be completely inactive in the skin. We reasoned that the contributions of viable populations to overall microbial abundance and functional community potential could be inferred by assessing bacterial growth rate from the skin metagenome. This would also allow us to estimate the ratio of rapidly growing to dead/stationary cells of common skin microbes, which would provide additional insights into antagonistic interspecies interactions.

To achieve this, we used the peak-to-trough ratio (PTR) method¹², implemented in GRiD¹³, which maps metagenomic reads to a bacterial reference genome to calculate coverage drops across the genome. Because most bacteria harbor a single circular chromosome replicated bi-directionally from the origin of replication (*ori*) to the terminus (*ter*) region¹⁴, rapidly growing cells will have a higher coverage at *ori* vs *ter*.

First, we generalized our analysis to define the steady-state growth dynamics of dominant skin microbes. Defining a microbe with a GRiD score > 1 as being in exponential phase, we identified the proportion of bacteria that were actively growing across different skin sites. Most examined microbes in dry sites (i.e., palm and forearm) were most active (Fig. 2A), interesting because these regions are often perturbed (e.g., by hand washing) and biomass is low. Such factors could affect microbial viability and growth rate to replenish the endogenous community. In contrast, sebaceous sites, which typically harbors higher biomass, least supported rapid growth, which could reflect the relatively specialized physiologic growth conditions, including anoxia, which would limit or slow growth of many microbes¹⁵. Foot sites favored growth of *S. epidermidis* and other non-lipophilic microbes at the expense of sebum-metabolizing *C. acnes*, which was least active at that site (Fig. 2A). Strikingly, growth rates of these microbes were stable over multiple timepoints (Fig. 2B).

We also asked if increased growth rate resulted in increased relative abundance. Relative abundance was positively correlated with growth rate only for certain species (Fig. 2C). For example, *C. ureicelerivorans*' relative abundance was strongly correlated with growth rate in all skin sites (Fig. 2C). This suggests that for select species, microbial abundance may be controlled by how rapidly cells are dividing. In contrast, *C. acnes* relative abundance and growth rate were strongly anti-correlated in sebaceous sites (Fig. 2C), suggesting that *C. acnes*' growth rate may

be modulated based on the presence/absence of competition or is regulated by its quorum sensing mechanisms.

Finally, we examined the microbial growth dynamics in patients with primary immunodeficiency syndrome, characterized with eczematous lesions⁵, as rapid growth might identify potential pathogens. Common skin commensals such as *S. epidermidis* and *C. ureicelerivorans* had significantly decreased growth rate in these patients (Fig. 2D). The decreased growth rate of *S. epidermidis* is surprising because these patients are unusually prone to staphylococcal infections. This observation may be due to the lack of correlation between growth rate and abundance for *S. epidermidis* in most skin sites (Fig. 2C).

Structural variants of *Staphylococcus epidermidis*

Community homeostasis is maintained by both host factors, interspecies interactions, and acquisition of genes, the latter of which can play important role in modulating adaptability to a particular environmental niche. These genes can encode for proteins that can influence signaling, virulence, and antimicrobial properties^{16,17}. Consequently, we examined microbial structural variants that may potentially harbor genes that play a role in strain adaptability to a specific skin site. We focused our analysis on *S. epidermidis* because unlike other *Staphylococci*, it thrives well in multiple skin sites (Fig. 2A).

We retrieved pangenome sequences from panDB, a database which assembles non-redundant genomic regions from multiple sequenced strains¹⁸, split sequences into 1 Kb fragments, and determined the enrichment of fragments across samples. Because *Staphylococcus* other than *S. epidermidis* thrives mostly in foot sites, and less in sebaceous regions (Fig. 2A), we investigated structural variants that might be associated with *S. epidermidis* adaptability in sebum-rich sites. We identified 6 *S. epidermidis* fragments that were always enriched in sebaceous sites when compared with other sites (Fig. 3A) which suggests that these may harbor genes essential for survival in sebaceous regions.

In addition, we hypothesized that if these fragments are indeed associated with adaptability, homologues in *S. capitis*, a closely related genome, will also be associated with adaptability in the latter. Similar to *S. epidermidis*, we identified 11 *S. capitis* fragments that were always differentially enriched in sebaceous sites (Fig. 3A). Interesting, 4 of the 7 genes harbored in *S. epidermidis* fragments had homologues in *S. capitis* (Fig. 3B), suggesting an underlying importance in sebaceous sites.

Next, we examined if these candidate genes could be located in mobile genetic elements which may suggest inter/intra- species transmission. Using BLASTn alignment to the NCBI nt database, all candidate genes could be identified in previously annotated plasmids or phages (Fig. 3C). We conducted additional benchmark analysis to minimize false positives by determining the correlation between candidate genes and *S. epidermidis* abundance or growth rate. Unsurprisingly, all fragments containing genes with homologues in *S. capitis* were positively correlated with both abundance and growth rate in sebaceous sites. Most of these genes encode for hypothetical proteins; however, a candidate (S_epi_13619) encodes for a stress response

protein suggesting that mechanisms to adapt to otherwise unfavorable conditions may play a important role in survival.

Indirect interactions: horizontally-transferred genes in skin microbes

To further investigate the genetic basis of interspecies interactions using our refined metagenomic analysis, we investigated horizontal gene transfer (HGT). HGT is a mechanism by which a microbe can acquire genetic material that may confer an increased survival or competitive advantage within a community.

We developed a HGT prediction pipeline (Supplementary Fig. 5) and based on the simulated dataset, the pipeline was able to identify 31% - 51% of the simulated HGT genes. We found that the prediction sensitivity was confined by the sensitivity of the metagenomic assembler (i.e., whether a gene was fully assembled) and the sensitivity of the gene predictor (i.e., whether an open reading frame was correctly annotated as a gene) (Supplementary Fig. 6), but not the synonymous distance-based algorithm presented in this study. The predicted HGT genes exhibited a large variation of synonymous distances, representing both recent HGT events and more ancient HGT events during the diversification of the microbial species (Supplementary Fig. 6). Importantly, the HGT genes with the lowest synonymous distances (synonymous distance < 0.1), which correspond to the most recent HGT events, almost exclusively matched the simulated HGT genes (94% - 100%, Supplementary Fig. 6), demonstrating the high specificity of the prediction pipeline.

To predict HGT among microbes within the skin microbiome, we developed a pipeline using our existing metagenomic data. In each pair of assembled genomes, we identified HGT candidates by looking for pairs of genes that are significantly more similar than immobile genes (Fig. 4A, Supplementary Fig. 5). Our HGT identification pipeline is a parametric version of a previously described robust method that searched for identical or near-identical gene pairs in distantly related microbial genomes¹⁹. Consistent with previous reports, functional annotation of HGT candidates showed a distribution over a wide functional spectrum¹⁹ (Fig. 4B). The most overrepresented functionality of the predicted HGT candidates were the transporters, highlighting the potential of the microbes to acquire the ability to uptake environmental nutrients and extrude harmful molecules through HGT events (Fig. 4B). Although most types of transporters were uniformly distributed in the mobile gene pool, microbiome at the sebaceous sites demonstrate enrichment of transporters that are involved in transporting metallic cation (Fig. 4B and 4C), including iron, manganese, zinc, cobalt, nickel, and biotin. These results suggest the existence of a pool of (conditionally) mobile genes exerting a multitude of biological functions, with enrichment of specific functions observed at specific skin types.

Finally, we constructed a network to reflect the top HGT events among microbial species in each type of skin sites (Figure 4D). Across skin sites, HGT were identified as a function of abundant species. For example, HGT candidates identified at a sebaceous site predominantly came from *Actinobacteria*, including bacteria in the genera *Propionibacterium*, *Corynebacterium* and *Staphylococcus* (Fig. 4D). Due to its dominance in many skin sites, *C. acnes* was central in networks corresponding to all types of skin types except for toenails, in which *Actinobacteria* (*C.*

acnes, *Corynebacterium singulare*, *Micrococcus luteus*, and *Kocuria rhizophila*) and multiple *Staphylococcus* species formed disconnected networks, strongly suggesting that HGT events at a body site were driven by the microbiome composition at the site. Overall, this pipeline characterizes statistically likely HGT candidates directly from shotgun metagenomic data, and is useful to simultaneously estimate the functional and taxonomic distribution of the mobile genes between genera.

DISCUSSION

The human skin, our largest organ and first line of defense, is home to a diversity of microorganisms. These microbes play an essential role in influencing metabolic processes, immune system modulation, and antagonism of potentially transient pathogens¹⁵. Alterations in community composition have been associated with a number of skin diseases like atopic dermatitis, psoriasis, and eczema^{20,21}. Numerous host-intrinsic (e.g., genetics, immune competence, skin barrier conditions), environmental or lifestyle factors (e.g., hygiene, exposure to different microbes), as well as microbiome-intrinsic factors affect disease predilection. A deep understanding of these factors, as well as the ecological complexity of the skin's microbiota, is needed to understand factors that influence its homeostasis and predisposition to disease.

Large-scale studies have aimed to characterize the skin microbiota using deep shotgun metagenomic sequencing, yet conclusions drawn from those previous analyses were limited in that a majority of sequence reads could not be mapped to any known genome^{5,6}. To solve this limitation, we used an integrated approach that incorporated *de novo* assembly and binning with reference-based analyses to identify and quantify new microbial skin occupants. This dual approach provided a key improvement in the resolution of the community. While we observed a significant reduction in uncharacterized sequence space using *de novo* approaches, viral genomes and low abundance genomes were poorly captured but could be resolved with reference genomes.

High-level analyses of the defining characteristics of the skin microbiome were largely concordant with previous findings. However, our hybrid approach provided important new insights into skin community structure, including: increased diversity, reduced stability across many sites, dominance of previously uncharacterized microbes in certain skin sites like the external auditory canal, and reduced representation of phage than previously believed. With few exceptions, we found that previous reports overestimated stability of the skin microbiome, likely because they lacked deeper classification of additional genomes, particularly for staphylococci and corneforms. Subsequently, we focused our analysis—continuing to interleave reference-based and *de novo* approaches—to examine potential factors that could underlie community stability and homeostasis, including the growth rate of different community members and interspecies interactions.

Leveraging metagenomic data to predict growth rate of dominant skin species, we measured marked variance in which species were actively growing in the skin, and how skin site could affect activity. For example, most microbes in dry sites were actively dividing compared to relatively few in sebaceous sites. Moreover, *S. epidermidis* appeared to grow exponentially in all sites, at similar

growth rates, suggesting that the skin environment generally provides adequate nutrients to support its rapid growth. Alternatively, *S. epidermidis* strains may have acquired specific genes modulating adaptability to each site. For example, we identified numerous genes present in mobile genetic elements that were associated with adaptability to sebaceous sites. Yet, strikingly, its ultimate relative abundance was not correlated with growth rate in most skin sites, suggesting an equally rapid killing or cell death. In this case, the lack of correlation between growth rate and relative abundance indicated that competitiveness within the community is largely independent of growth rate. Other skin sites/species showed a different relationship between growth rate and relative abundance. A positive correlation suggests that strains out-compete the rest of the community during exponential growth, as in the case of *C. ureicelerivorans*. Conversely, a negative correlation may reflect an active quorum sensing mechanism involved in the regulation of growth rate. For example, we observed a negative correlation for *C. acnes* that is restricted to sebaceous skin, suggesting that the microenvironment may play a role in regulating growth rate.

In addition, we observed that several types of transporters, especially the metal transporters, were highly abundant in the HGT gene pools at the sebaceous sites, potentially reflecting the importance to transport a vast variety of lipid soluble metals diffusing through the permeable cells of the sebaceous glands and follicular walls – a major absorption pathway of metals, including metallic toxicants. Additionally, the enrichment of metal transporters at the sebaceous sites paralleled the over-representation of *Actinobacteria* species at those skin sites, raising the possibility that the dissemination of metal transporting ability among *Actinobacteria* microbes at the sebaceous sites may be important to metal balance and consequently the health of the host.

In conclusion, we present a new landscape of the skin microbiome, providing the highest resolution reconstruction of microbial community composition and biodiversity in the skin to date. Importantly, we have used new approaches and analyses in a multifaceted investigation of functional elements and interspecies interactions underlying stability and community homeostasis. Our findings have generated testable hypotheses to interrogate interspecies and/or interstrain inhibition. Finally, our analyses are broadly applicable to investigate these mechanisms in skin disease, which ultimately can provide clues as to sources of antimicrobials directed towards strain-specific pathogens.

METHODS

Sample datasets

We retrieved 698 metagenomic shotgun skin samples from our previous work^{5,6}, which have been quality filtered for the presence of human DNA. The majority of the samples (n = 594) in these dataset were derived from longitudinal sampling of 12 individuals at 3 different time points with sampling intervals of 10-30 months (“long”) and 5-10 weeks (“short”). 24 samples from this collection were also derived from 2 individuals with hyper IgE syndrome. The remaining samples represent a single timepoint from three additional healthy individuals.

Taxonomic classification of skin microbes

To classify skin microbes, we used a hybrid *de novo*-based and referenced-based

characterization technique (Supplementary Fig. 2). All samples were concatenated and sequence reads were assembled into contigs using MEGAHIT 1.0.5 (--k-min 37 --k-max 67 --k-step 10 -m 0.99 --kmin-1pass --continue)¹¹, which was used for its ability to handle large datasets with relatively low memory requirement and short run time. We discarded contigs < 1Kb, mapped each individual sample's reads back to the contig catalog using bowtie2 2.2.8 (--sensitive)²², and extracted unmapped reads. However, read coverage from our derived contigs catalogue was relatively low (Supplementary Fig. 2). To re-assemble unmapped reads, we concatenated a subset of unmapped reads from randomly selected samples (due to the high memory requirement of SPAdes²³) and re-performed previous steps using SPAdes 3.7.1 (--meta) for *de novo* assembly, which was better able to resolve scaffolds from contigs. Newly extracted contigs/scaffolds were merged with the previous catalog to produce a non-redundant contigs catalog. We repeated the mapping of reads to the catalog to retrieve unmapped reads, randomly concatenated a small subset of unmapped reads, and performed assembly using SPAdes. We repeated this iterative step until no additional improvement in reads coverage from our contigs/scaffolds catalog was observed (Supplementary Fig. 2), obtaining a total of 1,037,465 contigs/scaffolds > 1Kb.

Contigs/scaffolds were then grouped into genome bins using MetaBAT (--sensitive, -m 1500)²⁴, which resulted in 556 bins. Bins were taxonomically classified using MEGAN 4.70.4²⁵ (Supplementary Table 1). We excluded 22 bins which were of non-microbial origin and further evaluated the quality of each bin using CheckM 1.0.6²⁶. For stringent annotation, we required that ≥ 65% of contigs/scaffolds present in a bin were assigned to the lowest level taxonomy; the sole exception being the kingdom-level taxonomy where our requirement was 40%. Bins were labeled "uncharacterized" if they were unable to be assigned to at least a kingdom, although by further relaxing our requirements to exclude contigs/scaffolds with no hits enabled characterization of those bins to at least the kingdom level (Supplementary Fig. 3B).

Each sample was then mapped back to the genome bins using bowtie2. Unmapped reads were subsequently characterized by a reference-based approach using reprDB¹⁸ and assigned to a genome using Pathoscope 2.0⁸. Reads unassignable by either method were categorized as "dark matter".

Bacterial growth rate estimation from skin samples

We estimated bacterial growth rate using GRiD (v1.3)¹³ using a coverage cutoff of 0.2. We created custom GRiD database using species of *C. acnes*, *S. epidermidis*, *S. aureus*, *C. jeikeium*, and *C. ureicelerivorans*.

Structural variant analysis

We retrieved pangenome sequences for *S. epidermidis* and *S. capitis* from panDB¹⁸, split genomes into 1 Kb fragments, and predicted genes using prokka²⁷. We mapped reads using bowtie2 to the fragment pool of each species, estimated reads count across samples using featureCounts²⁸, and DESeq2²⁹ to infer genes differentially enriched between skin sites. We filtered candidates using a adjusted p-value of 0.05 and log₂ fold change of 1. We aligned genes from *S. epidermidis* and *S. capitis* to generate a phylogenetic tree using MAFFT³⁰.

Horizontal gene transfer prediction

By definition, HGT genes that transferred between two lineages have shorter evolutionary distance than immobile genes (those that diverged at the time of divergence of the lineages). Therefore, HGT genes tend to be more similar¹⁹. For a pair of genomes, one reliable method of identifying HGT genes is to search for pairs of genes that are more similar to each other than the immobile genes in the genomes that reflect the evolutionary distance of the genomes. Genomes assembled from metagenomic shotgun data are inevitably incomplete and inaccurate, especially at strain level. Therefore, we assembled genomes at species level and proceeded only with genomes that have at least 25% completeness. To do this, metagenomic shotgun reads were first assembled using MEGAHIT. The resulting contigs were then assigned a taxonomic label using Kraken v0.10.6³¹; all contigs assigned to the same species were combined to represent the species draft genome. Genes were predicted from each species draft genome using prodigal³², and completeness of the assemblies assessed using BUSCO v2³³.

We identified potential HGT genes in each pair of species genomes. To do this, we first assembled a set of immobile genes from the genome pair to compute a null distribution of sequence similarity. Immobile genes were identified using the bacteria-specific universal single-copy orthologs (USCOs) annotated by BUSCO. Because USCOs are universally present across bacterial lineages and exist only in single copy, their horizontal transfer is unlikely. Second, all gene sequences from a pair of species genomes were clustered using uclust³⁴ at 0.5 similarity cut-off to reduce complexity. If any pair of genes within a cluster and from different species genomes has a significantly higher similarity than the immobile genes, the gene pair is identified as horizontally transferred. To mitigate the effect of natural selection, we computed synonymous distance—the number of synonymous changes per synonymous site—as the test statistics for similarity. Each pair of protein-coding genes was first reverse-aligned using the seqinr package³⁵, after which synonymous distance was computed using PAML³⁶, which implements an *ad hoc* method that corrects for codon frequency bias. Finally, to further lessen the influence of purifying selection, we removed HGT candidates that represented essential genes (i.e., genes with positive hits to the DEG 10 database³⁷ using ublast³⁴ with e-value < 10⁻⁹) and ribosome genes (ie, genes corresponding to KEGG BRITE ko03009 and ko03011).

We applied the prediction pipeline to three simulated microbial communities for validation. The simulated communities was generated using HgtSIM³⁸. Briefly, each community included three common skin bacteria species: *C. acnes*, *S. epidermidis*, and *Streptococcus mitis*. Each species was represented by three sequenced strain genomes, including one RefSeq representative strain and two other strains from RefSeq. All 9 strains were mixed at equal abundances in each simulated community. From the representative strain, 5 genes were randomly selected and horizontally transferred to all other strains in the community (a total of 15 HGT genes in each community). The three microbial communities differ in the amount of mutations accumulated in the HGT genes: the HGT genes were allowed to accumulate 0, 5% and 10% mutated bases in the recipient strains for the three communities, respectively. 3 million paired-end metagenomic shotgun reads were sampled from each community as part of the HgtSIM pipeline, and subsequently processed using the HGT prediction pipeline described above.

Statistics

We conducted all statistical analyses using R software. Spearman correlation was utilized for all correlation coefficient analyses while statistical differences between population groups were determined using the Wilcoxon rank-sum test. Where multiple measurements (e.g., timepoint, skin site within an individual) were used for correlation analyses, partial Spearman correlation was used, adjusting for multiple measurements. To assess microbial stability over time, we utilized the Yue-Clayton theta index, which calculates the distance between communities based on relative abundance of species in the population³⁹. Community diversity was determined using the Shannon diversity index, which measures both species richness and evenness.

DECLARATIONS

Ethics approval and consent to participate. Not applicable.

Consent for publication. All authors have approved submission of this manuscript.

Availability of data and material. The data used in this analysis is available in SRA under Bioproject 46333.

Competing interests. All authors declare that they have no competing interests.

Funding. This work was funded by the National Institute of Health (DP2 GM126893-01 and K22 AI119231-01). JO is additionally supported by the National Institutes of Health (1U54NS105539, 1 U19 AI142733, 1 R21 AR075174, 1 R43 AR073562), the Department of Defense (W81XWH1810229), the National Science Foundation (1853071), the American Cancer Society, and Leo Foundation.

Authors' contributions. AE and JO conceived the project. WZ contributed scripts and analysis. AE and JO analyzed data. AE and JO wrote the manuscript.

Acknowledgements. We would like to thank the Oh lab for critical reading of the manuscript.

REFERENCES

1. Nielsen, H. B. *et al.* Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat Biotechnol.* **32**, 822-828 (2014).
2. Cleary, B. *et al.* Detection of low-abundance bacterial strains in metagenomic datasets by eigengenome partitioning. *Nat Biotechnol.* **33**, 1053-1060 (2015).
3. Luo, C. *et al.* ConStrains identifies microbial strains in metagenomic datasets. *Nat Biotechnol.* **33**, 1045-1052 (2015).
4. Tsai, Y. C. *et al.* Resolving the Complexity of Human Skin Metagenomes Using Single-Molecule Sequencing. *MBio.* **7**, e01948-01915 (2016).
5. Oh, J. *et al.* Biogeography and individuality shape function in the human skin metagenome. *Nature.* **514**, 59-64 (2014).
6. Oh, J. *et al.* Temporal Stability of the Human Skin Microbiome. *Cell.* **165**, 854-866 (2016).
7. Rinke, C. *et al.* Insights into the phylogeny and coding potential of microbial dark matter.

- Nature*. **499**, 431-437 (2013).
8. Hong, C. *et al.* PathoScope 2.0: a complete computational framework for strain identification in environmental or clinical sequencing samples. *Microbiome*. **2**, 33 (2014).
9. Segata, N. *et al.* Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat Methods*. **9**, 811-814 (2012).
10. Tettelin, H. *et al.* Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proc Natl Acad Sci U S A*. **102**, 13950-13955 (2005).
11. Li, D., Liu, C. M., Luo, R., Sadakane, K., Lam, T. W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*. **31**, 1674-1676 (2015).
12. Korem, T. *et al.* Growth dynamics of gut microbiota in health and disease inferred from single metagenomic samples. *Science*. **349**, 1101-1106 (2015).
13. Emiola, A., Oh, J. High throughput *in situ* metagenomic measurement of bacterial replication at ultra-low sequencing coverage. *Nature Communications*. **9**, 4956 (2018).
14. Wang, J. D., Levin, P. A. Metabolism, cell growth and the bacterial cell cycle. *Nat Rev Microbiol*. **7**, 822-827 (2009).
15. Grice, E. A., Segre, J. A. The skin microbiome. *Nat Rev Microbiol*. **9**, 244-253 (2011).
16. Sharon, G. *et al.* Specialized metabolites from the microbiome in health and disease. *Cell Metab*. **20**, 719-730 (2014).
17. Donia, M.S., Fischbach, M.A. HUMAN MICROBIOTA. Small molecules from the human microbiota. *Science*. **349**, 1254766 (2015).
18. Zhou, W., Gay, N., Oh, J. ReprDB and panDB: minimalist databases with maximal microbial representation. *Microbiome*. **6**(1),15 (2018).
19. Brito, I. L. *et al.* Mobile genes in the human microbiome are structured from global to individual scales. *Nature*. **535**, 435-439 (2016).
20. Zeeuwen, P. L., Kleerebezem, M., Timmerman, H. M., Schalkwijk, J. Microbiome and skin diseases. *Curr Opin Allergy Clin Immunol*. **13**, 514-520 (2013).
21. Oh, J. *et al.* The altered landscape of the human skin microbiome in patients with primary immunodeficiencies. *Genome Res*. **23**, 2103-2114 (2013).
22. Langmead, B., Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. **9**, 357-359 (2012).
23. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol*. **19**, 455-477 (2012).
24. Kang, D. D., Froula, J., Egan, R., Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ*. **3**, e1165 (2015).
25. Huson, D. H., Auch, A. F., Qi, J., Schuster, S. C. MEGAN analysis of metagenomic data. *Genome Res*. **17**, 377-386 (2007).
26. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res*. **25**, 1043-1055 (2015).
27. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*. **30**, 2068-2069 (2014).

28. Liao, Y., Smyth, G.K., Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*. **30**, 923-930 (2013).
29. Love, M.I., Huber, W., Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*. **15**, p.550 (2014).
30. Katoh, K., Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. **30**, 772-780 (2013).
31. Wood, D.E., Salzberg, S.L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome biology* **15**(3) p.R46 (2014).
32. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*. **11**, 119 (2010).
33. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*. **31**, 3210-3212 (2015).
34. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*. **26**, 2460-2461 (2010).
35. Charif, D., Lobry, J. R. in *Structural Approaches to Sequence Evolution: Molecules, Networks, Populations* (Bastolla, U., Porto, M., Roman, H. E., Vendruscolo, M. eds.)(Springer, 2007).
36. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. **24**, 1586-1591 (2007).
37. Luo, H., Lin, Y., Gao, F., Zhang, C.T., Zhang, R. DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements. *Nucleic acids research* **42**(D1), pp.D574-D580 (2013).
38. Song, W., Steensen, K., Thomas, T. HgtSIM: a simulator for horizontal gene transfer (HGT) in microbial communities. *PeerJ* **5**, p.e4015 (2017).
39. Yue, J.C., Clayton, M.K. A similarity measure based on species proportions. *Commun. Stat. A-Theor*. **34**, 2123–2131 (2005).

FIGURE LEGENDS

Figure 1. Hybrid *de novo* and reference-based approach resolves dark matter in skin metagenome.

(A) Simplified flowchart of the hybrid *de novo* and reference-based characterization. **(B)** Boxplots show the fraction of uncharacterized sequences when using reference database from Oh *et al.*⁶, *de novo* genome bins only, or hybrid approach. Black lines indicate median; boxes first and third quartiles. **(C)** Microbial relative abundance across skin sites from a representative individual using the different classification approaches. **(D)** Heatmap shows microbial relative abundance across all samples from hybrid *de novo* and reference-based characterization, segregated by skin site characteristic. Darker colors indicate higher relative abundance. **(E)** Community diversity using Shannon diversity index. * = p-value < 0.05, ** = p-value < 0.01, *** = p-value < 0.001, NS = not significant by Wilcoxon-rank sum test. **(F)** Estimation of community stability using Yue-Clayton theta index, where $\theta \sim 1$ represents a completely stable community. “Long” refers to sampling time interval between T1 and T2 while “Short” represents short sampling time interval between T2 to T3. * = p-value < 0.05, ** = p-value < 0.01, *** = p-value < 0.001, NS = not significant by Wilcoxon-rank sum test.

Figure 2. Growth dynamics of skin bacteria.

Growth rate was calculated using GRiD¹³. **(A)** Proportion of common skin microbes in exponential vs. stationary growth phase/absent across different skin sites. GRiD = 1 for a microbe was considered stationary phase. *C. acnes*= *Cutibacterium acnes*, *C*=*Corynebacterium*, *S*=*Staphylococcus*. **(B)** Boxplot showing the GRiD score of microbes over time and grouped by skin site characteristics. **(C)** The heatmap shows partial Spearman correlation coefficient values, correcting for multiple measurements, between growth rate (GRiD) and relative abundance. Black colors indicate no significant correlation. The scatter plots below shows correlation between *C. acnes* growth rate (GRiD) and relative abundance from representative skin sites. **(D)** Boxplot showing microbial growth rate (GRiD) in healthy and primary immunodeficiency cohorts. Significant differences between groups were computed with Wilcoxon rank-sum test.

Figure 3. Structural variants in *Staphylococcus epidermidis*.

(A) Venn diagram showing number of fragments differentially enriched in sebaceous sites when compared with dry, moist, or foot sites. **(B)** Phylogenetic tree constructed for genes identified in candidate fragments of *S. epidermidis* and *S. capitis*. Genes from both species clustering together are highlighted. **(C)** Candidate *S. epidermidis* fragment, corresponding gene, and functional annotation. **(D)** Spearman correlation between candidate *S. epidermidis* gene fragments and relative abundance or growth rate (GRiD) in sebaceous and foot sites.

Figure 4. Horizontal transfer of genes in skin community.

(A) Overview of the HGT candidate identification process. Metagenomic reads were assembled and contigs belonging to the same species were pooled into a species draft genome. For each pair of species draft genomes, orthologous gene pairs were predicted. If a gene pair had a significantly smaller synonymous distance than the immobile gene pairs (that is, the universal single-copy orthologs), the pair of genes were identified as HGT candidates. **(B)** Distribution of functions (i.e., KEGG BRITES annotations) of all identified horizontal gene transfer (HGT) candidates and candidates that were annotated as transporters. **(C)** Detailed functions (i.e., KEGG orthologs) of HGT candidates annotated as metallic cation, iron-siderophore, and vitamin B12 transporters identified in the sebaceous sites. **(D)** Networks representing the top 10 species pairs for which HGT events were most frequent (i.e., HGT events were identified in the largest amount of samples) in each type of skin site. Nodes represent species and edges represent HGT events. In each type of skin site, the size of a node is proportional to the degree of that node.

SUPPLEMENTARY INFORMATION

Supplementary Figure 1. Skin sites, skin physiologic characteristics, and number of samples used. Overview of 698 samples analyzed in this study, encompassing 15 healthy adults and two hyper-IgE patients, three timepoints, and 17 skin sites, representing 4 microenvironments; dry, moist, sebaceous, and foot sites. The numbers adjacent each site correspond to the total number of samples derived from those sites.

Supplementary Figure 2. Flowchart of hybrid *de novo* and reference-based approach to

resolve microbial dark matter. The inset boxplot represents the percentage of reads mapping back to contigs/scaffolds (> 1 Kb) catalog after each round of iterative assembly. Black lines indicate median; boxes first and third quartiles.

Supplementary Figure 3. Hybrid *de novo* and reference-based resolution of dark matter.

(A) Scatterplots show concordance in relative abundance between the original classification in Oh *et al.*⁶ and the hybrid approach. Points deviating significantly from the diagonal are those whose relative abundance changed significantly based on the resolution of dark matter. Major genera, phyla, and kingdoms are shown. **(B)** Annotation of uncharacterized bins. Boxplots shows proportion of uncharacterized bins reassigned to a taxa. We relaxed our initial annotation requirement by excluding contigs/scaffolds with no hits from MEGAN output and re-ran our annotation pipeline.

Supplementary Figure 4. Higher compositional reconstruction of the skin microbiome.

(A) Community stability across skin site as calculated by the Yue-Clayton theta index, where $\theta \sim 1$ represents a completely stable community. “Long” refers to sampling time interval between T1 and T2 while “Short” represents short sampling time interval between T2 to T3. **(B)** Heatmap shows partial Spearman correlation values correcting for multiple measurements between microbial relative abundance at timepoints T2 vs T3 (i.e. short time interval) (top) and T1 vs T2 (i.e. long time interval) (bottom). Black colors indicate no correlation. *C. acnes*=*Cutibacterium acnes*, *C.*=*Corynebacterium*, *S.*=*Staphylococcus*.

Supplementary Figure 5. HGT candidate identification pipeline.

Supplementary Figure 6. Validation of the HGT prediction pipeline.

(A) Simulated HGT genes that were identified using the pipeline between all species pairs in all simulated communities (0%, 5%, and 10% mutations). **(B)** Number of HGT genes identified as a function of the synonymous distance of the HGT genes in the species pairs. Identified HGT genes that matched the simulated HGT genes were shown in red.

Supplementary Table 1: Microbial genome bins identified using *de novo* approach. Bins highlighted in yellow represent non-microbial genomes.

Supplementary Table 2: Bin coverage across samples.

Supplementary Table 3: Community relative abundance resolved using hybrid *de novo* and reference-based approach

A

B

C

D

E

F

A

B

C

D

A

B

C

Fragment	Gene	Functional annotation	
S_epi.fa_11535001-11536000	S_epi_13619	CsbD stress response protein	■
S_epi.fa_11535001-11536000	S_epi_13620	hypothetical protein	■
S_epi.fa_12070001-12071000	S_epi_14090	hypothetical protein	■
S_epi.fa_17193001-17194000	S_epi_19593	hypothetical protein	■
S_epi.fa_17193001-17194000	S_epi_19594	transcriptional regulator	■
S_epi.fa_17226001-17227000	S_epi_19640	hypothetical protein	■
S_epi.fa_18242001-18243000	S_epi_20627	hypothetical protein	■

■ Plasmid-borne
■ Phages

D

A

B

C

D

dark matter reads = total number of reads - (number of binned reads + number of Pathoscope-assigned reads)

A

B

Theta Index

1. All shotgun reads were assembled *de novo* and the resulting contigs belonging to the same species were pooled into a species draft genome.

2. Genes were predicted from assemblies of >25% completeness

3. Synonymous distances were computed between all pairs of genes from each pair of species. Additionally, a null distribution was computed using synonymous distances between all pairs of presumably immobile genes from each pair of species.

4. In each pair of species, gene pairs that have significantly smaller ($p < 0.05$) synonymous distances than the immobile genes were identified as HGT candidates.

A

B