


RESEARCH

Open Access



# A genome catalog of the early-life human skin microbiome

Zeyang Shen<sup>1</sup>, Lukian Robert<sup>1</sup>, Milan Stolpman<sup>2</sup>, You Che<sup>2</sup>, Katrina J. Allen<sup>3,4,5</sup>, Richard Saffery<sup>4,5</sup>, Audrey Walsh<sup>3,5</sup>, Angela Young<sup>3,5</sup>, Jana Eckert<sup>3,5</sup>, Clay Deming<sup>1</sup>, Qiong Chen<sup>1</sup>, Sean Conlan<sup>1</sup>, Karen Laky<sup>6</sup>, Jenny Min Li<sup>6</sup>, Lindsay Chatman<sup>6</sup>, Sara Saheb Kashaf<sup>1</sup>, NISC Comparative Sequencing Program, VITALITY team, Heidi H. Kong<sup>2</sup>, Pamela A. Frischmeyer-Guerrero<sup>6†</sup>, Kirsten P. Perrett<sup>3,4,5,7†</sup> and Julia A. Segre<sup>1\*†</sup> 

<sup>†</sup>Pamela A. Frischmeyer-Guerrero, Kirsten P. Perrett, and Julia A. Segre contributed equally to this work.

\*Correspondence: jsegre@nhgri.nih.gov

<sup>1</sup> Microbial Genomics Section, Translational and Functional Genomics Branch, National Human Genome Research Institute, NIH, Bethesda, MD, USA

<sup>2</sup> Dermatology Branch, National Institute of Arthritis and Musculoskeletal and Skin Diseases, NIH, Bethesda, MD, USA

<sup>3</sup> Population Allergy, Murdoch Children's Research Institute, Parkville, VIC, Australia

<sup>4</sup> Department of Paediatrics, University of Melbourne, Parkville, VIC, Australia

<sup>5</sup> Centre for Food and Allergy Research, Murdoch Children's Research Institute, Parkville, VIC, Australia

<sup>6</sup> Laboratory of Allergic Diseases, National Institute of Allergy and Infectious Diseases, Bethesda, MD, USA

<sup>7</sup> Department of Allergy and Immunology, Royal Children's Hospital, Parkville, VIC, Australia

## Abstract

**Background:** Metagenome-assembled genomes have greatly expanded the reference genomes for skin microbiome. However, the current reference genomes are largely based on samples from adults in North America and lack representation from infants and individuals from other continents.

**Results:** Here we use deep shotgun metagenomic sequencing to profile the skin microbiota of 215 infants at age 2–3 months and 12 months who are part of the VITALITY trial in Australia as well as 67 maternally matched samples. Based on the infant samples, we present the Early-Life Skin Genomes (ELSG) catalog, comprising 9483 prokaryotic genomes from 1056 species, 206 fungal genomes from 13 species, and 39 eukaryotic viral sequences. This genome catalog substantially expands the diversity of species previously known to comprise human skin microbiome and improves the classification rate of sequenced data by 21%. The protein catalog derived from these genomes provides insights into the functional elements such as defense mechanisms that distinguish early-life skin microbiome. We also find evidence for microbial sharing at the community, bacterial species, and strain levels between mothers and infants.

**Conclusions:** Overall, the ELSG catalog uncovers the skin microbiome of a previously underrepresented age group and population and provides a comprehensive view of human skin microbiome diversity, function, and development in early life.

## Background

In direct contact with the environment, human skin is both a barrier and a habitat for microbes, including bacteria, fungi, and viruses, which help modulate immune responses and provide colonization resistance from adverse species [1, 2]. Skin microbial community composition is shaped both by the ecology of the body site (oily, moist, dry) and skin physiology [1]. For example, during the transition through puberty, the



This is a U.S. Government work and not under copyright protection in the US; foreign copyright protection may apply 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

maturation of sebaceous glands creates a lipid-rich environment to facilitate growth of *Cutibacterium* [3]. Compared to adults, early-life skin is characterized by higher water content, lower natural moisturizing factor concentration, and fewer lipids [4, 5], which provides a distinct cutaneous environment for microbes and a unique habitat to study the skin microbiome.

Human skin microbiota is initially seeded at birth largely from maternal microbiome in association with the mode of delivery [6–8]. This relationship fades within 4–6 weeks [6, 7], but skin microbial communities at the species level were found to be similar between babies and mothers over weeks to years after delivery [6, 9, 10]. Even though multiple studies have investigated the transmission and development of the human gut microbiome [11–14], mother–infant transmission of skin microbiome remains underexplored. Specifically, microbial transmission on the skin has never been demonstrated at the resolution of strains.

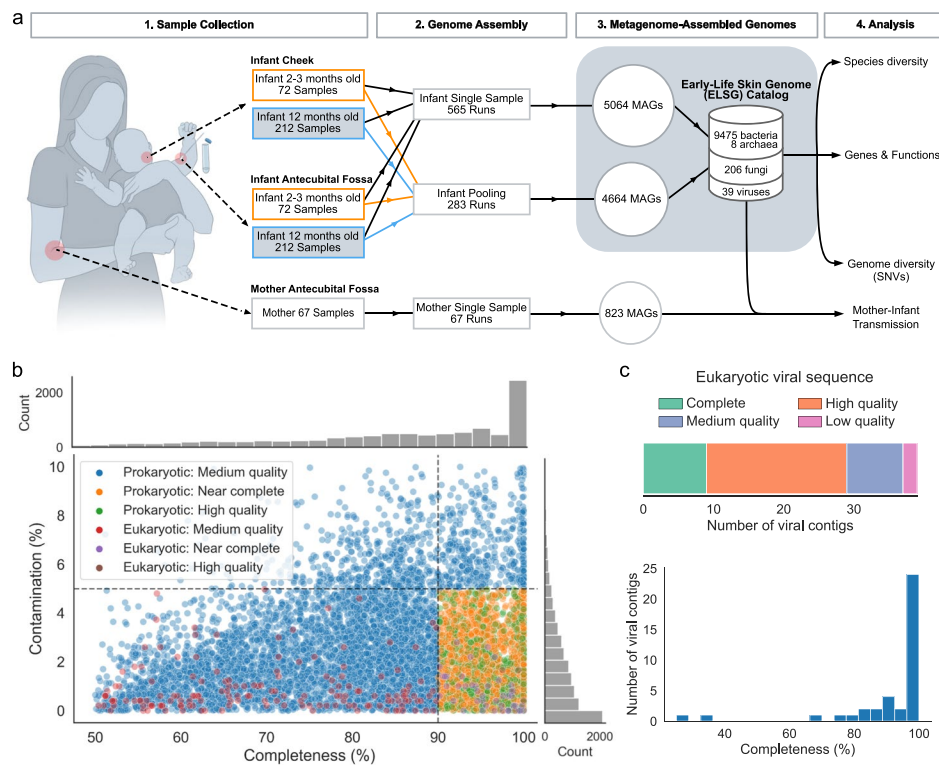
One major challenge in studying the early-life skin microbiome is the lack of microbial reference genomes. Previous skin metagenomic studies found approximately 50% of the metagenomic reads do not match genomes in public databases [1, 15]. Recent advancement in metagenome-assembled genomes (MAGs) has made it possible to generate large genome collections beyond culture-dependent methods [16]. We have recently published the Skin Microbial Genome Collection (SMGC) [17], which greatly expanded the reference genomes for skin microbiome in adults and substantially improved the classification rate of metagenomic reads. Comprehensive genome collections are also available for human gut microbiome [18–21]. In particular, the recent Early-Life Gut Genomes (ELGG) catalog has indicated great diversity and novelty of early-life gut microbiome compared to later in life [19]. To date, there have been no reports of skin microbial genomes in the first year of life. Comparative research investigating the gut microbiome in different populations also demonstrated great diversity of microbiome in people living in different geographic locations [18, 20, 21]. However, the current skin microbial genomes are derived from mostly adults residing in North America [17] and lack representation of individuals from other continents.

Here, we sequenced and assembled metagenomes from over 500 skin swabs collected longitudinally at age 2–3 months and 12 months from two body sites of 215 infants born in Australia, providing a catalog of 9728 genomes across multiple kingdoms for early-life skin microbiome. Using these data, we characterized the taxonomic and functional profile of the early-life skin microbiome and investigated the microbial sharing of the skin microbiome between mothers and infants.

## Results

### Deep sequencing of early-life skin metagenomes resulted in 9728 nonredundant microbial genomes

To obtain comprehensive skin microbiome on early life, we conducted deep shotgun metagenomic sequencing on 565 skin swabs collected from the cheek and antecubital fossa (inside bend of the elbow) of 215 infants who were part of the VITALITY trial [22] (Fig. 1a, Additional file 1: Table S1, S2). Among these infants, 69 were sampled longitudinally at 2–3 months and 12 months, 3 were sampled at 2–3 months only, and the rest were sampled at 12 months only. The two skin sites were selected as being



**Fig. 1** The genome catalog assembled from the early-life skin samples. **a** Schematic of study design from sampling to analysis. MAGs were constructed from single samples and pooled samples based on the two body sites of the same infant at each time point. MAGs from infant samples comprise the ELSG catalog. MAGs from mother samples were used for comparative analysis. **b** Completeness and contamination based on CheckM2 for each of nonredundant prokaryotic and eukaryotic MAGs included in the ELSG catalog, colored by the quality level. **c** Quality and completeness distribution for eukaryotic viral sequences included in the ELSG catalog

representative of sebaceous and moist sites, which are usually inhabited by distinct microbiomes [1] and have clinical importance for future eczema studies as these are commonly affected sites [23]. Each sample yielded a median of 28.6 million non-human reads (IQR = 11.7–48.6 million). We applied a previously established bioinformatic pipeline [24] to build MAGs from single samples. To increase MAG quality and the detection of rare species [17], we pooled reads from the two skin sites within the same individual at each time point to generate MAGs from an additional 283 co-assemblies (Fig. 1a). To generate MAGs, single and pooled samples were assembled with MEGAHIT [25] and binned with a combination of MetaBAT 2 [26], MaxBin 2 [27], and CONCOCT [28]. Prokaryotic MAGs were refined with metaWRAP [29] and checked for chimerism with GUNC [30], while eukaryotic MAGs were checked for quality with EukCC [31]. To elucidate viruses with a higher likelihood of affecting infants and causing infectious diseases, eukaryotic viral sequences were detected by aligning the contigs from MEGAHIT to the nucleotide collection database (nt) with BLASTn [32] and checked for quality with CheckV [33]. After removing redundant genomes across the entire dataset, our analyses yielded 9483 nonredundant prokaryotic MAGs, 206 nonredundant eukaryotic MAGs, and 39 eukaryotic viral sequences, comprising the Early-life Skin Genome (ELSG) catalog.

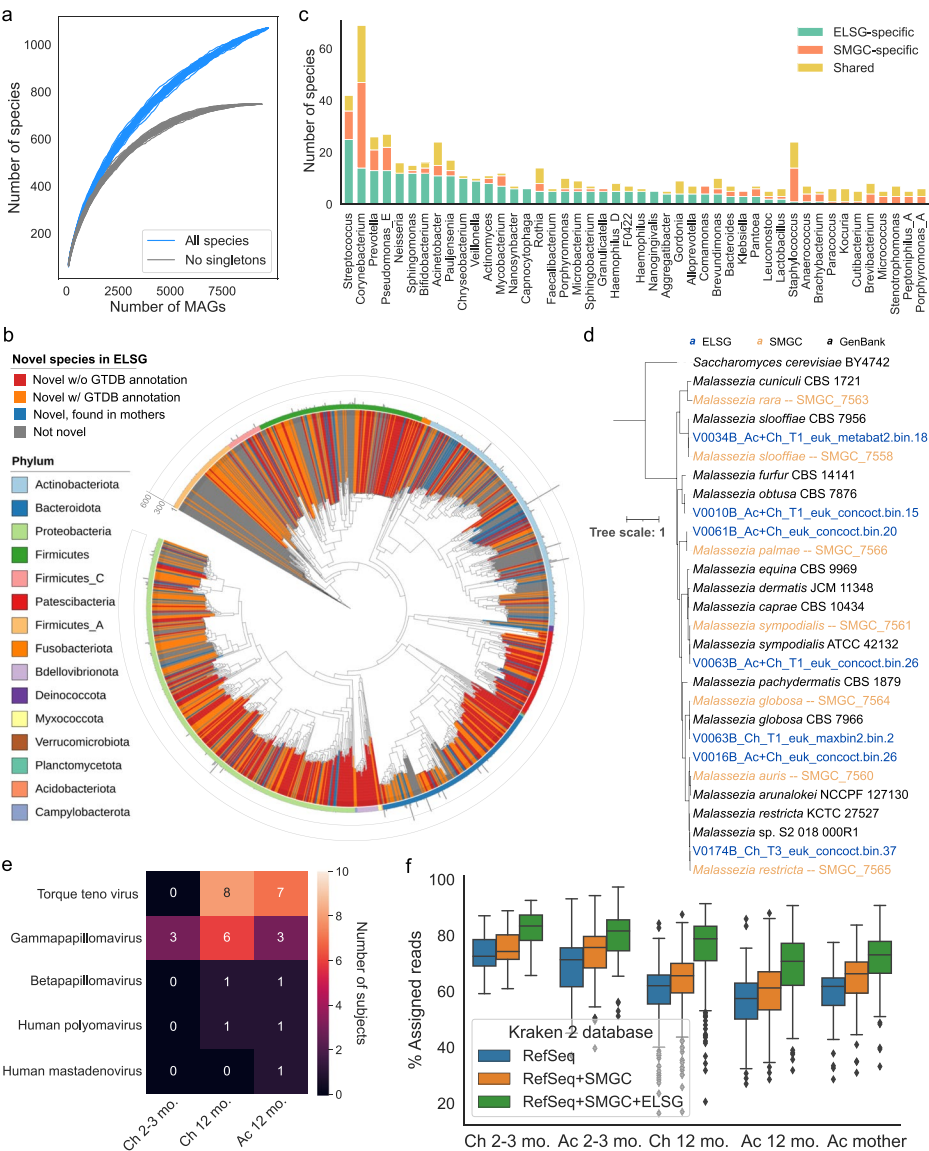
Among the 9483 prokaryotic MAGs (Fig. 1b, Additional file 2: Fig. S1a, Additional file 1: Table S3), 1578 were classified as “high-quality” (completeness > 90%, contamination < 5%, and the presence of 5S, 16S and 23S rRNA genes and at least 18 of the standard tRNAs); 2593 as “near-complete” (completeness > 90%, contamination < 5%, and did not meet the rRNA or tRNA requirement of high-quality MAGs); and 5312 as “medium-quality” (completeness > 50%, contamination < 10%, and quality score defined as completeness-5 × contamination [18] > 50) based on the Metagenome-Assembled Genome standard [34]. As a complement to the standard quality metrics, we estimated the level of strain heterogeneity of each MAG using CMSeq [16] and obtained the median at 0.17% for prokaryotic MAGs. We applied similar criteria to 206 eukaryotic MAGs, resulting in 5 “high-quality” (completeness > 90%, contamination < 5%, and the presence of 5S, 18S, 26S rRNA genes as well as at least 18 of the standard tRNAs), 42 “near-complete” (completeness > 90%, contamination < 5%, and did not meet the rRNA or tRNA requirement of high-quality MAGs), and 159 “medium-quality” MAGs (completeness > 50%, contamination < 10%) (Fig. 1b, Additional file 2: Fig. S1a, Additional file 1: Table S4). Higher quality of MAGs was usually associated with a lower number of contigs, a larger N50, a lower level of strain heterogeneity, a higher read depth, and the presence of more unique tRNAs (Additional file 2: Fig. S1a). Among the 39 eukaryotic viral sequences in the ELSG catalog, 9 were classified as “complete” (completeness = 100%), 20 as “high-quality” (completeness > 90%), 8 as “medium-quality” (completeness > 50%), and only 2 as “low-quality” (completeness < 50%) according to CheckV [33] (Fig. 1c, Additional file 2: Fig. S1b, Additional file 1: Table S5). Considering the challenge of assembling complete viral sequences from short-read metagenomes [33], we decided to include the two low-quality sequences in the ELSG catalog.

To compare the skin microbiome of infants with their mothers, we collected 67 skin swabs from the antecubital fossa of mothers during the 12-month infant visit (Fig. 1a). These samples underwent DNA sequencing and were assembled into individual sample-level MAGs using the aforementioned bioinformatic pipeline. The mother samples yielded a total of 764 bacterial MAGs, 1 archaeal MAG, 55 fungal MAGs, and 3 eukaryotic viral sequences of medium quality or higher.

### Species diversity in the ELSG catalog

To characterize the phylogenetic diversity of the ELSG catalog, we used 95% average nucleotide identity (ANI) threshold to further cluster the MAGs into 1055 bacterial, 1 archaeal, and 13 fungal species-level clusters [35]. Rarefaction analysis showed that the number of species in the ELSG was not saturated, when including MAGs recovered from a single sample. Excluding species recovered from only one sample, which may be transient in nature or individual-specific, the number of species came close to saturation, indicating that the ELSG catalog captured most of the common species present on the early-life skin (Fig. 2a).

Next, we explored the novelty of the species diversity in the ELSG catalog. We compared the ELSG catalog with the Skin Microbial Genome Collection (SMGC) [17], a collection of cultured and uncultured skin microbial genomes primarily based on adult samples in North America, and the Early-Life Gut Genome (ELGG) catalog [19]. Among the 1055 representative bacterial MAGs in the ELSG catalog, 743



**Fig. 2** Expansion of species diversity in skin microbiome. **a** Rarefaction analysis of the number of species as a function of the number of nonredundant genomes. Curves are depicted both for all the ELSG species and after excluding singleton species (represented by only one genome). **b** Phylogenetic tree of the representative bacterial MAGs in the ELSG catalog. Clades are colored by GTDB phylum annotation (outer ring) and whether these are novel species (inner shades). Bar graphs in the outermost layer indicate the number of nonredundant genomes within each species-level cluster. **c** Comparison of species diversity between the ELSG catalog and the SMGC. Species-level clusters were binned into the genus level in the bar graphs, ordered by a decreasing number of ELSG-specific species. **d** Phylogenetic tree of the *Malassezia* genomes from the ELSG and the SMGC together with GenBank reference genomes with *Saccharomyces cerevisiae* as the outgroup. **e** Number of infant samples harboring eukaryotic viruses included in the ELSG catalog. **f** Proportion of metagenomic reads from skin samples classified with Kraken 2 databases based upon RefSeq, augmented by the SMGC and the ELSG. The boxes represent the interquartile range, and the whiskers indicate the lowest and highest values within 1.5 times the interquartile range

clustered independently of any genome from the SMGC and the ELGG, expanding the phylogenetic diversity by 55% (Fig. 2b, Additional file 2: Fig. S2a, b, Additional file 1: Table S3). Among these, 339 were not assigned with species-level taxonomy



based on GTDB (red in Fig. 2b, Additional file 2: Fig. S2b). Note that 80 (11%) species-level clusters overlapped with MAGs built from mothers' skin samples (blue in Fig. 2b, Additional file 2: Fig. S2b), suggesting these species are likely population-specific rather than early-life-specific. ELSG-specific species spanned 15 different phyla greatly expanding the current knowledge of skin microbiome. Top genera of the early-life-specific species were *Streptococcus*, *Corynebacterium*, *Prevotella*, *Pseudomonas*, and *Neisseria* (Fig. 2c). Early-life species-level clusters that were also present in the SMGC-specific species were from the genera *Streptococcus*, *Corynebacterium*, and *Prevotella*. As some of the best studied skin genera, *Staphylococcus* harbored very few ELSG-specific species, and similarly, *Cutibacterium* species were almost always found in both the ELSG and the SMGC.

Among the eukaryotic genera covered by the ELSG catalog, *Malassezia* was the predominant genus, consistent with previous studies that found *Malassezia* being the major fungal genus across multiple skin sites [17, 36]. To compare the *Malassezia* species in the ELSG and the SMGC, we clustered 7 species-level representative MAGs from the ELSG classified to be *Malassezia*, 7 *Malassezia* MAGs from the SMGC, and representative GenBank reference genomes (Fig. 2d). Six species including *M. restricta*, *M. globosa*, *M. arunalokei*, *M. sympodialis*, *M. palmarum*, and *M. slooffiae* were shared by both the ELSG and the SMGC. Noticeably, *M. obtusa* was only assembled from the early-life skin but not found in the SMGC. Interestingly, *Saccharomyces* was the second largest fungal genus in the ELSG (Additional file 1: Table S4) but was not included in the SMGC or assembled from mother samples and was rarely studied in the context of skin. Together, these findings demonstrated the fungal diversity of early-life skin, as well as the commonalities and potential differences between early-life skin and adult skin.

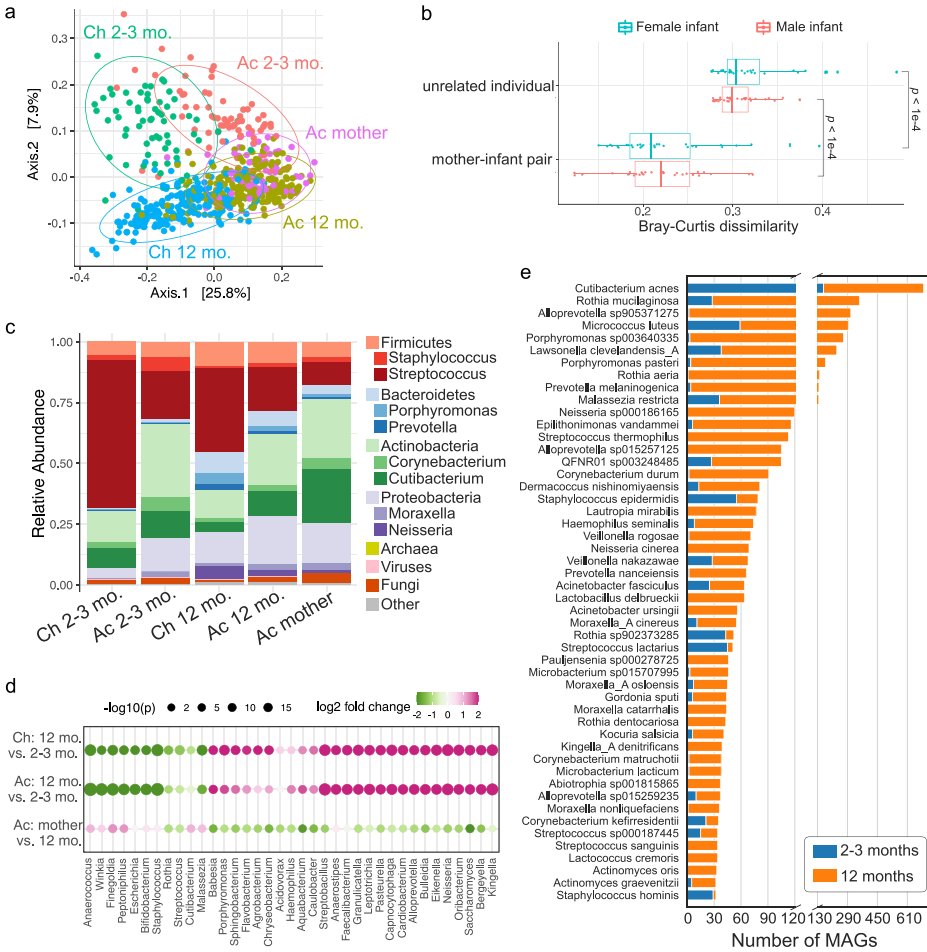
Next, we explored the species diversity of 39 eukaryotic viral sequences in the ELSG catalog. The most prevalent viruses found on infant skin were torque teno virus and gammapapillomavirus (Fig. 2e, Additional file 1: Table S5). Interestingly, the majority of these viral sequences were found exclusively in 12-month infants, except for the gammapapillomavirus discovered on the cheeks of three infants at 2–3 months.

Considering the novel species discovered on early-life skin, we used the ELSG catalog as an additional source of reference genomes to classify shotgun metagenomic reads. By adding the ELSG to a Kraken 2 database [37] created from the default RefSeq genomes and the SMGC, we obtained a median classification rate of 77% (IQR = 69–83%) for the early-life skin metagenomic datasets, which was a median of 21% improvement over the standard RefSeq database (Fig. 2f, Additional file 2: Fig. S2c). For the samples that did not directly contribute MAGs to the ELSG catalog, the median classification rate and improvement rate were 75% and 17%, respectively (Additional file 2: Fig. S2d). Interestingly, the ELSG also substantially improved the classification rate for metagenomic data of mothers (Fig. 2f, Additional file 2: Fig. S2c) and slightly improved read mapping for the antecubital metagenomes of the SMGC (Additional file 2: Fig. S2e), suggesting the value of the ELSG in capturing age- or population-specific species.

### Comparison of taxonomic profiles between early-life and adult skin microbiome

We next explored similarities of the infant skin microbial community at two time points as well as the relatedness of infant skin to mothers. The microbial community

of infants demonstrated strong skin-site differentiation with cheek and antecubital samples separated on a principal coordinate analysis as well as age differentiation with 2–3 months and 12 months separated for each skin site (Fig. 3a). Interestingly, the microbial community on the antecubital fossa of mothers was most similar to the antecubital fossa of infants at 12 months (Fig. 3a), suggesting a potential trajectory of maturation in the microbial community from early life to adulthood. We calculated Bray–Curtis dissimilarity between the antecubital fossa of babies and mothers and saw a significantly lower beta diversity ( $p < 1e-4$ ) between related infant–mother pairs compared to unrelated infant–mother pairs, consistent for both infant sexes (Fig. 3b). We also calculated the beta diversity between the two time points of the same infant



**Fig. 3** Early-life skin microbial community structure. **a** Principal coordinate analysis (PCoA) on Bray–Curtis dissimilarity between the microbial profiles. Each point represents a single sample and is colored by body site and age group. Ellipses represent a 95% confidence interval around the centroid of each sample group. **b** The Bray–Curtis dissimilarity of mother–infant pairs comparing related versus unrelated dyads. Median value of each infant and all unrelated mothers was used. Statistical difference was tested by two-sided Wilcoxon rank sum test. **c** Relative abundance of skin microbiome averaged for each sample group. Two of the most abundant genera within each bacterial phylum were shown. **d** Differential taxa at the genus level between infants of different ages and between infants at 12 months and mothers. The size of the dots represents the log-transformed adjusted p-value from DESeq2, and the color indicates fold changes. The top differentially abundant genera for each comparison were shown. **e** Number of species-level MAGs recovered from infants at 2–3 months and 12 months, sorted by the total number of MAGs

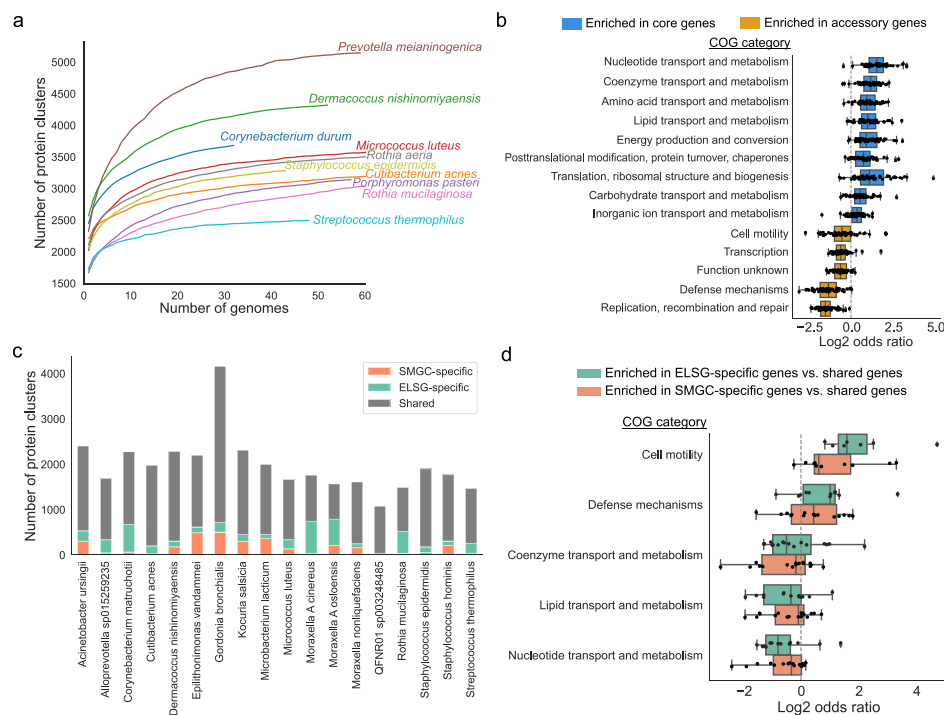
as compared to different individuals. For both body sites, we saw a significantly lower beta diversity ( $p < 0.01$ ) within the same individuals, indicating an individualized trajectory of maturation that starts as early as 2–3 months (Additional file 2: Fig. S3a). Together, this suggests that the microbial communities on infant skin may be influenced by individual factors, including the mother's skin microbiome.

Overall, the skin microbiome of early life contained roughly 97.3% bacteria, 2.4% fungi, and 0.3% viruses (Fig. 3c) or 92% bacteria, 1% fungi, and 7% viruses after genome size normalization (Additional file 2: Fig. S3b). Antecubital fossa of infants generally had a more diverse microbial community than the cheek (Additional file 2: Fig. S3c). We also saw an increase in diversity from 2–3 months to 12 months at both body sites, with the richness of 12-month-old antecubital fossa resembling that of mothers (Additional file 2: Fig. S3c). At the phylum level, Actinobacteria were more abundant on the antecubital fossa of both infants and mothers, whereas more Firmicutes, particularly *Streptococcus*, was found on cheek (Fig. 3c). Infants gained Bacteroidetes and Proteobacteria on both skin sites over time (Fig. 3c). While displaying the most similar profile to the 12-month-old antecubital fossa, the maternal antecubital fossa exhibited higher Actinobacteria and reduced Firmicutes (Fig. 3c). Differential abundance analysis indicated 165 genera significantly (adjusted  $p < 0.01$ ) gained abundance at antecubital fossa over time and 209 genera increased on the cheek, including *Neisseria* and *Saccharomyces* (Fig. 3d). Another 69 genera and 55 genera lost abundance at 12 months on antecubital fossa and cheek, respectively, including *Staphylococcus* (Fig. 3d), which is consistent with previous studies that also found a decrease in *Staphylococcus* over time [7, 38]. When compared to maternal skin, 12-month-old infants showed significantly decreased *Malassezia* and increased *Saccharomyces*, aligning with the fungal diversity observed in the ELSG. The prevalence of abundant species was correlated with the number of genomes in the ELSG (Additional file 2: Fig. S3d). For instance, *Cutibacterium acnes* was the most prevalent species found on early-life skin and contributed the largest number of MAGs in the ELSG (Fig. 3e). Consistent with a higher abundance of *Staphylococcus* at 2–3 months, most of the *Staphylococcus* genomes were assembled from infants at 2–3 months even though the sample size at 2–3 months is much smaller than 12 months (Fig. 3e).

#### Comparison of the early-life and adult skin microbiome protein catalogs

To estimate the functional capacity in the ELSG catalog, we predicted protein-coding sequences for each of the 9483 prokaryotic MAGs, resulting in a total of ~3.5 million protein clusters at 90% amino acid identity. According to the rarefaction analysis, the protein clusters found in the ELSG catalog were not saturated, but close to saturation when only considering ~2 million protein clusters that were identified in at least two MAGs (Additional file 2: Fig. S4a), consistent with previous findings in gut microbiome [18, 19]. When examining individual species, we discovered that some of the most prominently represented species had either reached a saturation point or were nearing saturation (Fig. 4a). The conspecific gene frequency had a bimodal distribution (Additional file 2: Fig. S4b), consistent with observations in the SMGC [17]. We defined those genes shared by at least 90% conspecific genomes of each species as core genes and the rest as accessory genes [18] (Additional file 2: Fig. S4c) and then compared the functions encoded in the core and accessory genes based on several annotation databases.

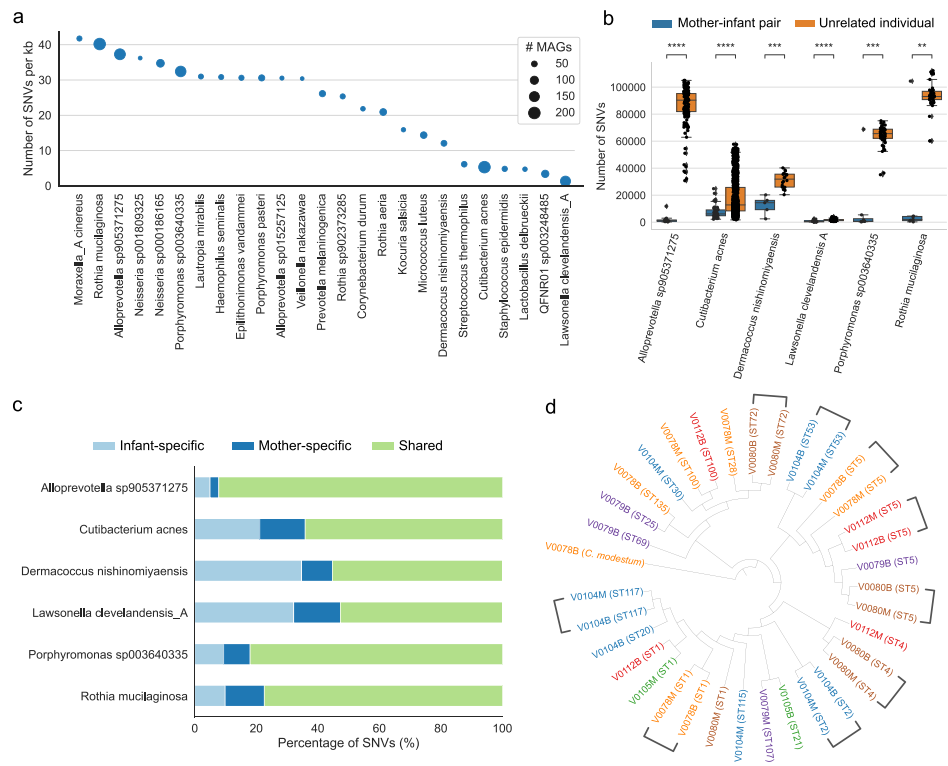




**Fig. 4** Proteins and functions of early-life skin microbiome. **a** Rarefaction curves of the number of protein clusters obtained as a function of the number of species-level genomes. Each curve represents one species. The curves for species with more than 60 genomes are truncated for visualization purpose. **b** Comparison of the functional categories assigned to the core and accessory genes for species with at least 10 near-complete or high-quality genomes (> 90% completeness, < 5% contamination). Each dot represents one species. Odds ratio was calculated from the contingency table with core and accessory genes on one axis and the tested and the other functional categories on the other axis. Only significantly enriched functional categories are shown. Significance was calculated with a two-tailed *t*-test on log-transformed odds ratios and further adjusted for multiple comparisons using the Bonferroni correction. **c** Comparison of the protein clusters between the ELSG and the SMGC for species with at least five near-complete or high-quality genomes in each catalog. **d** Functional categories enriched in ELSG-specific and SMGC-specific genes compared to shared genes. Each dot represents a species. Only statistically significant categories are shown

Core genes were generally better annotated than accessory genes in all databases (Additional file 2: Fig. S4d). According to COG annotations [39], core genes were enriched for functions related to metabolism and translation, whereas accessory genes were enriched for functions related to replication, defense mechanisms, and transcription (Fig. 4b). A similar pattern of functional roles performed by core and accessory genes has previously been reported for gut microbiomes [18].

We next compared the pan-genome of early-life skin microbiome with that of SMGC. The pan-genome size was variable between the two genome collections for several species (Additional file 2: Fig. S4e). For example, *Micrococcus luteus* had a 14% larger pan-genome in the ELSG catalog, while, in contrast, *Cutibacterium acnes* had a 5% larger pan-genome in the SMGC. Besides the pan-genome size difference, many genes were specific to one collection (Fig. 4c). Interestingly, ELSG- or SMGC-specific genes were enriched in COG categories such as cell motility and defense mechanisms while collection-shared genes were enriched for functions related to metabolism (Fig. 4d).



**Fig. 5** Single-nucleotide variation indicates mother–infant microbial sharing. **a** Top species with the largest intraspecies SNV density. The size of dots indicates the number of MAGs corresponding to each species. **b** Number of SNVs in pairwise comparisons between mother–infant pairs and between infants and unrelated mothers. Only species with genomes from at least four mother–infant pairs were considered for analysis. Statistical significance was tested by two-tailed Wilcoxon rank sum test. \*\* $P < 0.01$ , \*\*\* $P < 0.001$ , \*\*\*\* $P < 0.0001$ . **c** Proportion of SNVs that were found in genomes from infants only or mothers only or both. SNVs were called based on the species-level representative MAG as the reference genome. **d** Phylogenetic tree of representative *C. acnes* cultured isolates with *C. modestum* as the outgroup. Source of individual is indicated in the label name and label color. Sequence type is displayed in parentheses

**Intraspecies genomic diversity indicates microbial sharing between infants and mothers**

To characterize the genomic diversity across species-level clusters within the ELSG catalog, we calculated the rate of intraspecies single-nucleotide variants (SNVs). *Rothia mucilaginosa*, a prevalent species on early-life skin, contained one of the highest SNV density, 40 SNVs per kb, suggesting a great potential of functional variability (Fig. 5a). By contrast, *Cutibacterium acnes*, which was even more prevalent, had a much lower density of only about 5 SNVs per kb. Similarly, *Staphylococcus epidermidis*, another common species found on skin, had about 5 SNVs per kb.

Next, we compared paired microbial genomes from infants and mothers. For all six species for which we had MAGs from at least four related infant–mother pairs, there were significantly fewer SNVs genome-wide ( $p < 0.01$ ) between related infant–mother pairs as compared to unrelated infants and mothers, potentially due to the vertical transmission of skin microbes between mothers and infants (Fig. 5b). By looking at SNVs at protein-coding regions, three of the six species including *Cutibacterium acnes* had 62% or less SNVs shared by infants and mothers, whereas the other three species including *Rothia mucilaginosa* had over 78% of SNVs shared by infants and mothers (Fig. 5c). The small proportion of age–group-specific SNVs within these

three species was also consistent with the strikingly large differences between related and unrelated babies and mothers (Fig. 5b).

Besides the genome sharing between infants and mothers, we also investigated the genome sharing at different ages of infants. For the five species with at least four infants that yielded longitudinal pairs of MAGs, the number of SNVs was generally lower within individuals than across individuals (Additional file 2: Fig. S5a), suggesting temporally persistent microbial genomes on the host. Due to a limited number of samples, further research is needed to examine the applicability of such observation to a broad spectrum of species.

To further validate the mother–infant microbial sharing, we cultured *Cutibacterium acnes* from the nasal swabs collected from six pairs of infants and mothers when infants were 12 months old. Nares harbor a greater microbial biomass compared to the skin and engage in frequent microbiome exchange with the skin, making them a suitable proxy for inferring strain sharing of the skin microbiome. Depending on the variable viability of bacteria, we were able to obtain and sequence 4–12 *C. acnes* independent colonies from each individual (Additional file 1: Table S6). Genomes from the related infants and mothers were often closely placed on a phylogenetic tree (Fig. 5d). Consistent with that, we performed multi-locus sequence typing to these genomes and found that four out of six mother–infant pairs shared at least one sequence type (Additional file 2: Fig. S5b), which is statistically significant ( $p = 0.012$ ) based on a permutation test (Additional file 2: Fig. S5c). Together, this provided evidence for the mother–infant microbial sharing at the strain level and established the basis for future research on dissecting the transmission pathways of such sharing.

## Discussion

We present the first genome collection for early-life skin microbiome and the largest skin microbial genome collection to date containing over a thousand species-level clusters of bacterial and fungal genomes and an additional set of eukaryotic viral sequences. To our knowledge, the ELSG catalog is also the first skin microbial genome collection based on samples from Australia. Nevertheless, the geographic specificity of our data prompts a thoughtful consideration of the extent to which our conclusions can be extended to encompass diverse infant populations worldwide. The validation of the ELSG as an effective resource of improving read classification for infants from distinct geographic backgrounds is an avenue for future research. The slightly improved classification of North American samples by including the ELSG catalog could be due to the deep sequencing and the large sample basis of this study, which recovered ultra-rare and low-abundant species present on human skin across continents. Augmented read mapping would be consistent with species that are more abundant in infants and at lower abundance in adults. The ELSG catalog includes hundreds of species previously not characterized for skin, many of which are novel species. Considering that skin is still an understudied organ source of microbiome, this study has demonstrated the importance of profiling different age groups and populations to capture a complete catalog of human skin microbiome. Since the ELSG catalog was based on infant samples at age 12 months or less, this resource will be of particular use in studies of childhood cutaneous disorders, such as atopic dermatitis, which commonly begins in infancy.

Our study on the skin microbial sharing was empowered by a substantial number of paired samples collected from infants and mothers. Evidence of skin microbial sharing was found at various levels, encompassing the microbial community, individual species, and even strains. Specifically, infants and their mothers had closely related microbial profiles, relatively similar conspecific MAGs, and shared strains of *Cutibacterium acnes*. Furthermore, our longitudinal assessment, which involved infants at 2–3 months and 12 months, revealed signs of temporal persistence within the infant skin microbiome. This persistence was evident in both microbial profiles and genomes. These findings underscore the significant influence mothers exert in shaping the skin microbiome during early life and suggest a potential impact of preceding states on later microbiome compositions. However, it is important to note that our study does not exclude the possible contribution of other sources such as fathers or environments to the observed mother–infant microbial sharing. This is evident as two out of the six mother–infant pairs where we cultured *C. acnes* isolates shared none of their *C. acnes* strains. Thus, a comprehensive understanding of the microbial transmission pathways and directions between mothers, infants, and other potential sources requires further investigation. Subsequent studies should also endeavor to expand our findings to encompass other species not investigated within the scope of this study.

Based on the ELSG catalog, we analyzed the largest published protein catalog for skin microbiome to estimate the functional capacity. By looking at the conspecific pan-genomes, we summarized the functional categories that distinguish core and accessory genes, which replicated the findings in gut microbiome. Interestingly, genes found only in one of the two current skin genome collections were consistently represented by functions related to defense mechanism and replication, recombination, and repair. These categories are potentially the drivers of functional specificity in early-life skin microbiome. Further experiments are needed to validate the function and importance of individual genes in maintaining homeostasis on early-life skin.

## Conclusions

In summary, our investigation involved profiling the skin metagenomes of infants who had been previously under-represented. This pioneering effort led to the development of the ELSG catalog, which significantly expands the repertoire of skin microbial genomes in infants. The ELSG catalog presents a comprehensive and versatile resource for future studies focused on various aspects of the infant skin microbiome such as microbial transmission and development, and the intricate interplay between disease and the early-life skin microbiota.

## Methods

### Participant recruitment, skin sampling, and metagenomic sequencing

New mothers along with their infants were recruited as part of the VITALITY trial [22]. Written informed consent was obtained for all participants in this study. Skin samples were collected from the antecubital fossa and cheek of 72 infants at ages 2–3 months. Sixty-nine of these infants together with 140 additional infants were sampled at the same sites at age 12 months. In addition, 67 of these infants' mothers were sampled at the antecubital fossa during the same visit when the 12-month samples were taken. To

maximize microbial recovery, no bathing was permitted within 24 h of sample collection. Skin was sampled with an established protocol using pre-moistened Puritan foam swabs collected and stored in 100  $\mu$ L Yeast Cell Lysis Buffer (Lucigen) buffer at  $-80^{\circ}$  and shipped on dry ice. Concomitant with skin sample collection, air swabs were collected as negative controls to account for any potential environmental or reagent contaminants.

Samples were converted to genomic DNA with an established protocol [40, 41]. Briefly, DNA libraries for Illumina sequencing were prepared using the Nextera XT DNA Library Preparation Kit (Illumina) per manufacturer's instructions with the exception of increasing the AMPure XP Bead clean-up volume from 30  $\mu$ L to 50  $\mu$ L; 1 ng of extracted DNA was used as input into the fragmentation step. DNA is simultaneously fragmented and tagged with sequencing adapters in a single-tube enzymatic reaction. Libraries were then sequenced with the Illumina NovaSeq 6000 sequencing platform at the NIH Intramural Sequencing Center for  $2 \times 150$  bp, 50 million paired-end reads per sample.

Most of the negative controls yielded  $<1\%$  of the reads derived from skin samples except for one. We excluded the skin samples collected at the same time of that air swab together with one infant's antecubital fossa sample which yielded less than 10,000 reads. Our final set of samples for analysis includes 565 from infants (424 at 12 months (212 infants  $\times$  2 skin sites) + 144 at 2–3 months (72 infants  $\times$  2 skin sites)—3 samples failed) and 67 from mothers.

#### Bacteria culturing and sequence typing

Nasal culture samples were obtained from infants and mothers during the same visit when infants were 12 months old using the COPAN eSwab system in 1 mL AMIES and frozen at  $-80^{\circ}$  C. Broths were diluted and plated on Brain Heart Infusion Agar (BHI + 10  $\mu$ g/mL Fosfomycin) and incubated in an anaerobic chamber for 7 days at  $37^{\circ}$  C. Colonies were screened with PCR using *C. acnes*-specific primers PA-1 5'-GGG TTGTAAACCGCTTTCGCTG-3 and PA-2 5'-GGCACACCCATCTCTGAGCAC-3, then streaked for purity on Blood Agar plates (TSA with 5% Sheep Blood – Remel R01201). gDNA was prepared from isolates and sequenced with an established protocol [17]. *C. acnes* genomes were assembled from sequenced reads using SPAdes [42] and checked for quality using the "lineage\_wf" workflow of CheckM v1.1.3 [43]. The sequence type of each *C. acnes* genome was identified by multi-locus sequence typing scheme from PubMLST [44]. *C. acnes* genomes of the same individual were first dereplicated at 99.9% ANI with dRep v3.2.2 [45] and then used to build the phylogenetic tree with GToTree v1.6.37 [46] based on the single-copy gene set of Actinobacteria.

#### Pre-processing, metagenomic assembly, and contig binning

Metagenomic reads were trimmed for adapters with Cutadapt v3.4 using the parameters "`-nextseq-trim 20 -e 0.15 -m 50`" [47] and checked for quality with PRINSEQ-lite v0.20.4 using the parameters "`-lc_method entropy -lc_threshold 70 -min_len 50 -min_qual_mean 20 -ns_max_n 5 -min_gc 10 -max_gc 90`" [48]. Reads with less than 50 bp length after trimming were removed. The reads were then aligned to the GRCh38 human reference genome with Bowtie2 v2.4.5 using the parameters "`-very-sensitive`" [49]. The human reads were removed before assembly.



Metagenomic assembly was performed with MEGAHIT v1.2.9 using the default parameters [25]. Pool individual runs were conducted after concatenating the reads from the two skin sites of the same infant at each time point. We performed 283 co-assemblies including 211 from 12 months and 72 from 2–3 months. Contigs were then binned with a combination of MetaBAT 2 v2.15 [26], MaxBin 2 v2.2.7 [27], and CONCOCT v1.1.0 [28] by running the binning module of metaWRAP v1.3.2 [29] with the parameter “-l 1500” indicating the minimal contig length 1500 bp.

### Genome quality assessment

To obtain prokaryotic MAGs, the bins produced by each binning tool were refined with the Bin\_refinement module of metaWRAP v1.3.2 [29]. The completeness and contamination of refined bins were evaluated with the “lineage\_wf” workflow of CheckM v1.1.3 [43] and the “predict” function of CheckM2 v1.0.2 [50]. The quality score was calculated as: completeness – 5 × contamination. Ribosomal RNAs in each genome were detected with the “cmsearch” function of INFERNAL v1.1.4 using parameters “-anytrunc -noali” [51] against the Rfam covariance models for the 5S (5S\_rRNA), 16S (SSU\_rRNA\_bacteria), and 23S rRNAs (LSU\_rRNA\_bacteria) [52]. Transfer RNAs of the standard 20 amino acids were identified with tRNAScan-SE v2.0.11 using the parameter “-B” for bacterial species [53]. Each genome was assessed for chimerism with GUNC v1.0.5 [30]. The MAGs with contamination greater than 0.05, clade separation greater than 0.45 and a reference representation score greater than 0.5 were excluded. Based on the Metagenome-Assembled Genome standard [34], MAGs with >90% completeness, <5% contamination, the presence of 5S, 16S, and 23S rRNA genes, and at least 18 tRNAs were reported as high-quality draft genomes. MAGs with >90% completeness and <5% contamination but missing the rRNAs or tRNAs were reported as near-complete genomes. MAGs with >50% completeness, <10% contamination, and quality score >50 were reported as medium quality. We used an inclusive approach by considering all medium-quality MAGs satisfying the completeness, contamination, and quality score requirements based on at least one of CheckM and CheckM2, while the high-quality and near-complete MAGs were identified based on CheckM2 statistics.

To assess eukaryotic MAGs, the bins from the three binning tools were estimated for completeness and contamination with EukCC v2.1.0 [31]. rRNAs and tRNAs were identified using the same approach above except that the Rfam [52] covariance models 5\_S\_rRNA, SSU\_rRNA\_eukarya, and LSU\_rRNA\_eukarya were used to find 5S, 18S, and 26S, respectively. Bins with >90% completeness, <5% contamination, the presence of 5S, 18S, and 26S rRNA genes, and at least 18 tRNAs were reported as high-quality draft genomes. Those with >90% completeness and <5% contamination but not satisfying the rRNAs and tRNAs requirements were defined as near-complete. The remaining bins with >50% completeness and <10% contamination were reported as medium-quality genomes.

We further mapped each contig of MAGs to the nt database with BLASTn v2.8.0 [32] to assess viral contamination. Contigs with the top hit of a eukaryotic viral genome with >95% nucleotide identity, >1000 bp aligned sequence, and >70% total contig aligned were removed before quality assessment. The contig number and N50 of MAGs were

calculated using in-house scripts. Read depth was calculated by first mapping the raw reads back to MAGs Bowtie2 v2.4.5 [49] using the default parameters and then calculating mean depth with SAMtools v1.16.1 [54]. The strain heterogeneity was estimated by the “polymut.py” script of CMSeq v1.0.4 with parameters “-mincov 10 -minqual 30 -dominant\_frq\_thrsh 0.8” [16].

Eukaryotic viral contigs detected by the method above was further assessed with CheckV v1.0.0 based on CheckV database v1.5 [33] for completeness and contamination.

### Redundancy removal and species clustering

To remove redundant genomes that were recovered by both single and pooled sample runs, we dereplicated MAGs at a 99.9% ANI threshold with dRep v3.2.2 using parameters “-pa 0.999 -SkipSecondary -comp 50 -con 10” [45]. Dereplication was performed on prokaryotic MAGs and eukaryotic MAGs separately.

The MAGs were clustered at the species level by dereplicating at a 95% ANI threshold with dRep v3.2.2 using parameters “-pa 0.90 -sa 0.95 -nc 0.30 -cm larger -S\_algorithm fastANI -comp 50 -con 10 -run\_tertiary\_clustering -clusterAlg single” [45]. fastANI v1.33 [35] was used to accelerate the process. CheckM2 statistics were used as inputs. Representative genome of each species-level cluster was selected based on the dRep scores derived from genome completeness, contamination, strain heterogeneity, and contig N50.

### Taxonomic assignment and phylogenetic analysis

Taxonomic annotation of prokaryotic MAGs was assigned with the “classify\_wf” workflow of GTDB-Tk v2.1.0 using default parameters and GTDB database release 207 [55, 56]. The phylogenetic tree of bacterial representative genomes of species-level clusters was built with IQ-TREE v1.6.12 using the parameter “-m MFP” [57] based on the protein sequence alignments generated by GTDB-Tk.

The eukaryotic MAGs were compared with all of the GenBank fungal genomes first using Mash v2.3 [58] and then assigned species-level taxonomy with at least 95% ANI calculated by fastANI v1.33 [35]. The phylogenetic tree was built with the script BUSCO\_phylogenomics.py ([https://github.com/jamiecmg/BUSCO\\_phylogenomics](https://github.com/jamiecmg/BUSCO_phylogenomics)) based on single-copy marker genes identified by BUSCO v4.1.3 using the parameter “-m geno -f -auto-lineage-euk” [59]. The phylogenetic trees were visualized with iTOL [60]. The taxonomic classifications of viral sequences were assigned by the top alignment hit from BLASTn [32].

### Metagenomic read classification and microbial abundance estimation

Metagenomic reads were mapped with Kraken v2.1.2 using parameters “-confidence 0.1 -paired” [37] against the standard RefSeq database (release 211) and two custom database with additional representative genomes from the SMGC and ELSC catalogs. To integrate the genome catalogs with the RefSeq genomes, we first converted GTDB taxonomy to NCBI taxonomy using the “gtdb\_to\_ncbi\_majority\_vote.py” script available in the GTDB-Tk repository [56] and then obtained NCBI taxonomy IDs corresponding to the species- and genus-level taxonomy of each genome with taxonkit v0.12.0 [61]. We excluded 22 and 106 representative MAGs from the SMGC and the ELSC,

respectively, which did not have a match ID at the genus level. For MAGs with a match ID at the genus level but not at the species level, we created a new taxonomy ID associated with each MAG when building the Kraken databases. Classification improvement was calculated on a per-sample basis as  $(\text{proportion of reads classified with custom database} - \text{proportion of reads classified with RefSeq database}) / \text{proportion of reads classified with RefSeq database} \times 100$ . Species-level microbial abundances were computed with Bracken v2.5 using parameters “-r 100 -l S” [62].

#### Alpha and beta diversity calculation

Skin metagenomic data with less than 800,000 classified reads were excluded (4% of samples). The remaining samples were first rarefied and then calculated for the number of species with  $\geq 5$  reads (richness) and Shannon index with the “diversity” function of vegan package in R v4.1. To calculate the beta diversity, we first removed taxa present in  $\leq 20\%$  samples and then performed log transformation on species abundances after adding pseudocount 1. Bray–Curtis dissimilarity was calculated with the “distance” function of phyloseq v1.38.0 [63] in R v4.1. Principal coordinate analysis was conducted based on Bray–Curtis dissimilarity with the “ordination” function of phyloseq package.

#### Differential abundance analysis

Differential abundance was calculated with DESeq2 v1.34.0 [64] using the parameters “test = “Wald”, sfType = “poscounts”, fitType = “local”” based on the rarefied raw reads as used for diversity calculation. Low-prevalence taxa present in less than 10% of samples were removed. Comparisons were conducted for each of the two skin sites, comparing 2–3 months and 12 months and for antecubital fossa, comparing infants at 12 months and mothers. Significantly differential taxa were identified by  $< 0.01$  adjusted  $p$ -value and  $> 2$ -fold change.

#### Pan-genome analysis and functional annotation

Protein-coding sequences (CDS) of each genome were predicted and annotated with Prokka v1.14.6 using parameter “-kingdom Bacteria” [65]. Protein clustering across all species of with the “easy-linclust” function of MMseqs2 v13-45111 using parameters “-cov-mode 1 -c 0.8 -kmer-per-seq 80 -min-seq-id 0.9” to generate protein clusters at 90% amino acid identity, respectively [66].

The pan-genome analysis was performed only on near-complete and high-quality genomes. Species with at least ten near-complete or high-quality nonredundant genomes were analyzed with Panaroo v1.3.0 using the parameters “-clean-mode strict -merge\_paralogs -c 0.90 -core\_threshold 0.90 -f 0.5” for  $\geq 90\%$  amino acid identity and a family threshold of 50% [67]. Functional annotation of all protein sequences was performed with eggNOG-mapper v2.1.6 [68] to obtain COG [39], KEGG [69], Pfam [70], and GO [71] annotations.

#### SNV analysis

To assess SNV density of species, we first mapped conspecific genomes to the representative genome using the “nucmer” program of MUMmer v3.1 [72], filtered alignments

with the “delta-filter” program using parameters “-q -r”, and then identified SNVs with the “show-snps” program. SNV density of each genome was computed by dividing the number of SNVs by the size of the representative genome. Only SNVs which occurred in at least two conspecific genomes were included in the analysis. The final SNV density of each species was the mean of SNV densities of all conspecific genomes. The minimum number of conspecific genomes for SNV analysis was ten. The same programs and parameters were used for mother–infant genome comparisons.

### Statistical analysis

Statistical analyses were performed using *ggpubr* package in R v4.1 or *scipy* package in Python v3.9.9. Two-sided Wilcoxon rank sum tests and *t*-tests were used to evaluate differences between groups. Pearson correlation coefficient was used to assess correlation. Functional enrichment analysis was performed using two-sided Fisher’s exact test, with *p*-values adjusted by the Bonferroni method. The permutation test ( $n = 1,000$ ) was applied to assess the significance of sequence type sharing between mothers and infants.

### Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13059-023-03090-w>.

**Additional file 1: Supplementary Table 1.** Participant characteristics. **Supplementary Table 2.** Summary of skin metagenomes. **Supplementary Table 3.** Quality and taxonomic information of the prokaryotic MAGs. **Supplementary Table 4.** Quality and taxonomic information of the fungal MAGs. **Supplementary Table 5.** Quality and taxonomic information of the eukaryotic viral sequences. **Supplementary Table 6.** Summary of *Cutibacterium acnes* isolates.

**Additional file 2: Figure S1.** Quality metrics of the nonredundant genomes in the ELSG catalog. **Figure S2.** Expansion of species diversity in the ELSG catalog. **Figure S3.** Early-life skin microbial community structure. **Figure S4.** Proteins and functions of early-life skin microbiome. **Figure S5.** Intraspecies single-nucleotide variation and mother–infant strain sharing.

**Additional file 3.** Review history.

### Acknowledgements

The study made use of the computational resources provided by the NIH HPC Biowulf Cluster (<http://hpc.nih.gov>). The authors are grateful to the VITALITY families for participation.

### Consortia

#### NISC Comparative Sequencing Program

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

#### VITALITY team

Investigators: Prof Kirsten Perrett, Prof Katrina J. Allen, A/Prof Justin Brown, A/Prof Natalie Carvalho, Prof Nigel Curtis, Prof Kim Daziel, Prof Shyamali Dharmage, A/Prof Ronda Greaves, Prof Lyle Gurrin, Dr Li Huang, A/Prof Jennifer Koplin, Prof Katherine Lee, A/Prof Georgia Paxton, A/Prof Rachel Peters, Prof Anne-Louise Ponsonby, Dr Peter Sayre, Prof Mimi Tang, Prof Peter Vuillermin, Prof Melissa Wake.

Study Team: Dr Angela Young, Deborah Anderson, Christine Axelrad, Anna Bourke, Kirsty Bowes, Dr Tim Brettig, Natasha Burgess, Beatriz Camesella-Perez, Xueyuan Che, Daniela Ciciulla, Jac Cushnahan, Helen Czech, Dr Thanh Dang, Kathryn Dawes, Dr Jana Eckert, Hannah Elborough, Dr Michael Field, Charlie Fink, Sarah Fowler, Grace Gell, Rebecca Gray, Emi Habgood, Richard Hall, Phoebe Harris, Erin Hill, Kensuke Hoashi, Hannah Ilhan, Narelle Jenkins, Andrew Knox, Clare Morrison, Dr Melanie Neeland, Jenn Ness, Dr Wendy Norton, Sasha Odoi, Dr Mary Panjari, Kayla Parker, Ahelee Rahman, Ashleigh Rak, Maisie Ralphsmith, Natalie Schreurs, Carrie Service, Dr Victoria Soriano, Judith Spotswood, Dr Mark Taranto, Leone Thiele, Kate Wall, Audrey Walsh, Angela Walsh, Anita Wise

Data and Safety Monitoring Team: Prof Andrew Davidson, Prof Arul Earnest, Dr Lara Ford, Dr Andrew Kemp, Dr Sam Mehr, Dr Tibor Schuster, Dr Dean Tey, Diana Zannino.

Pharmacy Team: Donna Legge, Jason Bell, Joanne Cheah, Kay Hynes, Kee Lim, Emily Porrello, Annette Powell

Biobanking Team: Pedro Ramos, Anushka Karunanayake, Izabelle Mezzetti, Kayla Parker, Ronita Singh

Media and Communications: Harriet Edmund, Bridie Byrne, Tom Keeble, Cuby Martis, Belle Ngien

Legal: Penny Glenn, Andrew Kaynes

**Review history**

The review history is available as Additional file 3.

**Peer review information**

Wenjing She was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

**Authors' contributions**

JAS, KPP, and PAF conceived the project. ZS, YC, SC, and SSK performed the analyses. ZS, LR, and YC produced the figures. ZS and JAS wrote the manuscript. LR, MS, YC, KJA, RS, AW, AY, JE, CD, QC, SC, KL, JML, LC, and HHK reviewed and revised the manuscript. LR, MS, CD, QC, KL, JML, and LC performed the experiments. KJA, RS, AW, AY, and JE managed and curated the data. KPP and PAF supervised data collection. JAS and HHK supervised the experiments. NISC Comparative Sequencing Program sequenced the samples. VITALITY team collected and provided the samples. All authors read and approved the final manuscript.

**Funding**

Open Access funding provided by the National Institutes of Health (NIH) Research reported in this publication was performed in part as a project of the Immune Tolerance Network and supported by the National Institute of Allergy and Infectious Diseases (NIAID) of the National Institutes of Health (NIH) under Award Number UM1AI109565. This research received support from the NIH Intramural Research Programs of the National Human Genome Research Institute (NHGRI), the NIAID as well as the National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIAMS). The Vitality trial received support from the National Health and Medical Research Council (NHMRC) of Australia (GNT1146913); Epworth Medical Foundation, The Kimberley Foundation, DHB Foundation, Rotary Club of Camberwell, The Isabel & John Gilbertson Charitable Trust and individual donors.

**Availability of data and materials**

The raw metagenomic sequencing data are available in the NCBI BioProject database under project number PRJNA971252 [73]. The MAGs of the ELSG catalog can be found at <https://research.nhgri.nih.gov/projects/ELSG/> [74], where users have access to download nonredundant genomes, species-level representative genomes, phylogenetic tree files, protein catalog, pan-genome annotations, and a custom Kraken 2 database based on the ELSG catalog. All the code utilized in this study is available on GitHub: <https://github.com/skinmicrobiome/ELSG> [75] and Zenodo: <https://doi.org/10.5281/zenodo.8422805> [76]. Both repositories are released under the MIT license.

Other publicly available data used in this project: SMGC [17] is available at [http://ftp.ebi.ac.uk/pub/databases/metagenomics/genome\\_sets/skin\\_microbiome](http://ftp.ebi.ac.uk/pub/databases/metagenomics/genome_sets/skin_microbiome) [77]. Shotgun metagenomic sequencing data used in the SMGC is accessed from the NCBI Sequence Read Archive under accession number SRP002480 [78]. ELGG [19] catalog is available at <https://doi.org/10.5281/zenodo.6969520> [79].

**Declarations****Ethics approval and consent to participate**

The Vitality Trial is sponsored by Murdoch Children's Research Institute (<http://www.clinicaltrials.gov/ct2/show/NCT02112734>). This study was approved by the Royal Children's Hospital Human Research Ethics Committee (#34168) and conducted in accordance with the Declaration of Helsinki. Written informed consent was obtained for all subjects participating in this study.

**Competing interests**

The authors declare that they have no competing interests.

Received: 23 May 2023 Accepted: 17 October 2023

Published online: 10 November 2023

**References**

- Oh J, et al. Biogeography and individuality shape function in the human skin metagenome. *Nature*. 2014;514:59–64.
- Harris-Tryon TA, Grice EA. Microbiota and maintenance of skin barrier function. *Science*. 2022;1979(376):940–5.
- Park J, et al. Shifts in the skin bacterial and fungal communities of healthy children transitioning through puberty. *J Invest Dermatol*. 2022;142:212–9.
- Casterline BW, Paller AS. Early development of the skin microbiome: therapeutic opportunities. *Pediatr Res*. 2021;90:731–7.
- Stamatas GN, Nikolovski J, Mack MC, Kollias N. Infant skin physiology and development during the first years of life: a review of recent findings based on in vivo studies. *Int J Cosmet Sci*. 2011;33:17–24.
- Chu DM, et al. Maturation of the infant microbiome community structure and function across multiple body sites and in relation to mode of delivery. *Nat Med*. 2017;23:314–26.
- Capone KA, Dowd SE, Stamatas GN, Nikolovski J. Diversity of the human skin microbiome early in life. *J Invest Dermatol*. 2011;131:2026–32.
- Dominguez-Bello MG, et al. Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proc Natl Acad Sci*. 2010;107:11971–5.
- Bogaert D, et al. Mother-to-infant microbiota transmission and infant microbiota development across multiple body sites. *Cell Host Microbe*. 2023;31:447–460.e6.



10. Zhu T, et al. Age and mothers: potent influences of children's skin microbiota. *J Invest Dermatol*. 2019;139:2497–2505.e6.
11. Ferretti P, et al. Mother-to-infant microbial transmission from different body sites shapes the developing infant gut microbiome. *Cell Host Microbe*. 2018;24:133–145.e5.
12. Yassour M, et al. Strain-level analysis of mother-to-child bacterial transmission during the first few months of life. *Cell Host Microbe*. 2018;24:146–154.e4.
13. Valles-Colomer M, et al. The person-to-person transmission landscape of the gut and oral microbiomes. *Nature*. 2023. <https://doi.org/10.1038/s41586-022-05620-1>.
14. Valles-Colomer M, et al. Variation and transmission of the human gut microbiota across multiple familial generations. *Nat Microbiol*. 2022;7:87–96.
15. Oh J, Byrd AL, Park M, Kong HH, Segre JA. Temporal stability of the human skin microbiome. *Cell*. 2016;165:854–66.
16. Pasolli E, et al. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell*. 2019;176:649–662.e20.
17. SahebKashaf S, et al. Integrating cultivation and metagenomics for a multi-kingdom view of skin microbiome diversity and functions. *Nat Microbiol*. 2021;7:169–79.
18. Almeida A, et al. A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat Biotechnol*. 2021;39:105–14.
19. Zeng S, et al. A compendium of 32,277 metagenome-assembled genomes and over 80 million genes from the early-life human gut microbiome. *Nat Commun*. 2022;13:5139.
20. Jin H, et al. A high-quality genome compendium of the human gut microbiome of Inner Mongolians. *Nat Microbiol*. 2023;8:150–61.
21. Kim CY, et al. Human reference gut microbiome catalog including newly assembled genomes from under-represented Asian metagenomes. *Genome Med*. 2021;13:134.
22. Allen KJ, et al. VITALITY trial: protocol for a randomised controlled trial to establish the role of postnatal vitamin D supplementation in infant immune health. *BMJ Open*. 2015;5:e009377.
23. Kennedy EA, et al. Skin microbiome before development of atopic dermatitis: early colonization with commensal staphylococci at 2 months is associated with a lower risk of atopic dermatitis at 1 year. *J Allergy Clin Immunol*. 2017;139:166–72.
24. SahebKashaf S, Almeida A, Segre JA, Finn RD. Recovering prokaryotic genomes from host-associated, short-read shotgun metagenomic sequencing data. *Nat Protoc*. 2021;16:2520–41. <https://doi.org/10.1038/s41596-021-00508-2>.
25. 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*. 2015;31:1674–6.
26. Kang DD, et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*. 2019;7:e7359.
27. Wu Y-W, Simmons BA, Singer SW. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*. 2016;32:605–7.
28. Alneberg J, et al. Binning metagenomic contigs by coverage and composition. *Nat Methods*. 2014;11:1144–6.
29. Uritskiy GV, DiRuggiero J, Taylor J. MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome*. 2018;6:158.
30. Orakov A, et al. GUNC: detection of chimerism and contamination in prokaryotic genomes. *Genome Biol*. 2021;22:178.
31. Saary P, Mitchell AL, Finn RD. Estimating the quality of eukaryotic genomes recovered from metagenomic analysis with EukCC. *Genome Biol*. 2020;21:244.
32. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990;215:403–10.
33. Nayfach S, et al. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat Biotechnol*. 2021;39:578–85.
34. Bowers RM, et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat Biotechnol*. 2017;35:725–31.
35. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun*. 2018;9:5114.
36. Findley K, et al. Topographic diversity of fungal and bacterial communities in human skin. *Nature*. 2013;498:367–70.
37. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. *Genome Biol*. 2019;20:257.
38. Rapin A, et al. The skin microbiome in the first year of life and its association with atopic dermatitis. *Allergy*. 2023;78:1949–63.
39. Galperin MY, Makarova KS, Wolf YI, Koonin EV. Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res*. 2015;43:D261–9.
40. Tirosh O, et al. Expanded skin virome in DOCK8-deficient patients. *Nat Med*. 2018;24:1815–21.
41. Byrd AL, et al. Staphylococcus aureus and Staphylococcus epidermidis strain diversity underlying pediatric atopic dermatitis. *Sci Transl Med*. 2017;9:eaal4651.
42. Bankevich A, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol*. 2012;19:455–77.
43. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res*. 2015;25:1043–55.
44. Jolley KA, Bray JE, Maiden MCJ. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications [version 1; referees: 2 approved]. *Wellcome Open Res*. 2018;3.
45. Olm MR, Brown CT, Brooks B, Banfield JF. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J*. 2017;11:2864–8.
46. Lee MD. GToTree: a user-friendly workflow for phylogenomics. *Bioinformatics*. 2019;35:4162–4.
47. Martin M. Cutadapt removes sequences from high-throughput sequencing reads. *EMBnet J*. 2013;17:1.
48. Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics*. 2011;27:863–4.

49. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 2012;9:357–9.
50. Chklovski A, Parks DH, Woodcroft BJ, Tyson GW. CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nat Methods*. 2023;20:1203–12.
51. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*. 2013;29:2933–5.
52. Kalvari I, et al. Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Res*. 2021;49:D192–200.
53. Chan PP, Lin BY, Mak AJ, Lowe TM. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Res*. 2021;49:9077–96.
54. Danecek P, et al. Twelve years of SAMtools and BCFtools. *Gigascience*. 2021;10:giab008.
55. Parks DH, et al. A complete domain-to-species taxonomy for bacteria and archaea. *Nat Biotechnol*. 2020;38:1079–86.
56. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics*. 2022;38:5315–6.
57. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol*. 2015;32:268–74.
58. Ondov BD, et al. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol*. 2016;17:132.
59. Manni M, Berkeley MR, Seppely M, Simão FA, Zdobnov EM. BUSCO Update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol Biol Evol*. 2021;38:4647–54.
60. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res*. 2021;49:W293–6.
61. Shen W, Ren H. TaxonKit: a practical and efficient NCBI taxonomy toolkit. *J Genet Genomics*. 2021;48:844–50.
62. Lu J, Breitwieser FP, Thielen P, Salzberg SL. Bracken: estimating species abundance in metagenomics data. *PeerJ Comput Sci*. 2017;3:e104.
63. McMurdie PJ, Holmes S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One*. 2013;8:e61217.
64. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15:550.
65. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*. 2014;30:2068–9.
66. Steinegger M, Söding J. Clustering huge protein sequence sets in linear time. *Nat Commun*. 2018;9:2542.
67. Tonkin-Hill G, et al. Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol*. 2020;21:180.
68. Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol Biol Evol*. 2021;38:5825–9.
69. Kanehisa M, et al. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res*. 2014;42:D199–205.
70. Finn RD, et al. Pfam: the protein families database. *Nucleic Acids Res*. 2014;42:D222–30.
71. Consortium, T. G. O. Gene Ontology Consortium: going forward. *Nucleic Acids Res*. 2015;43:D1049–56.
72. Kurtz S, et al. Versatile and open software for comparing large genomes. *Genome Biol*. 2004;5:R12.
73. Shen Z, Frischmeyer-Guerrero P, Perrett K, Segre J. ELSG metagenome sequencing. Datasets. Sequence Read Archive. 2023. <https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA971252>.
74. Shen Z, Frischmeyer-Guerrero P, Perrett K, Segre J. ELSG catalog. Datasets. NHGRI; 2023. <https://research.nhgri.nih.gov/projects/ELSG/>.
75. Shen Z. ELSG source codes. GitHub; 2023. <https://github.com/skinmicrobiome/ELSG>.
76. Shen Z. ELSG - Early-Life Skin Genome Catalog. Zenodo; 2023. <https://doi.org/10.5281/zenodo.8422805>.
77. Kashaf SS, Segre JA, Almeida A, Finn RD. SMGC. Datasets. EBI; 2021. [https://ftp.ebi.ac.uk/pub/databases/metagenomics/genome\\_sets/skin\\_microbiome/](https://ftp.ebi.ac.uk/pub/databases/metagenomics/genome_sets/skin_microbiome/).
78. Segre JA. Skin metagenome sequencing. Datasets. Sequence Read Archive. 2021 <https://www.ncbi.nlm.nih.gov/sra/?term=SRP002480>.
79. Zeng S et al. Early-life human gut metagenome-assembled genomes and proteins catalogs. Zenodo; 2022. <https://zenodo.org/doi/10.5281/zenodo.6969519>.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.