

A global genomic and structural resource for the human skin microbiome

Zhiming Li

lzm405@126.com

BGI-Shenzhen <https://orcid.org/0000-0001-5589-1136>

Yanmei Ju

BGI-Shenzhen <https://orcid.org/0000-0002-1600-3164>

Yijing Zhan

Chinese Academy of Medical Sciences and Peking Union Medical College

Bo Zhang

Chinese Academy of Medical Sciences and Peking Union Medical College

Shenghui Li

Puensum Genetech Institute <https://orcid.org/0000-0003-1071-8510>

Shutian Lin

BGI Precision Nutrition

Ting Qiu

BGI Precision Nutrition <https://orcid.org/0009-0003-4846-7419>

Hefu Zhen

BGI-Shenzhen, Shenzhen 518083

Shunxiang Jiang

BGI Precision Nutrition

Xinlan Zhou

BGI Research, Shenzhen 518083, China

Jingjing Xia

IUF-Leibniz Research Institute for Environmental Medicine

Shaoqiao Zhang

BGI-Research, Wuhan, 430074, China <https://orcid.org/0009-0009-3430-4161>

Jiucun Wang

State Key Laboratory of Genetic Engineering, Collaborative Innovation Center for Genetics and Development, School of Life Sciences and Human Phenome Institute <https://orcid.org/0000-0003-2765-0620>

Qianmei Liu

Chinese Academy of Medical Sciences and Peking Union Medical College

Tao Zhang

BGI-Shenzhen <https://orcid.org/0000-0003-2765-2802>

Yuanqiang Zou

BGI Research <https://orcid.org/0000-0003-2380-264X>

Xin Jin

BGI Research <https://orcid.org/0000-0001-7554-4975>

Wenwei Zhang

BGI-Shenzhen <https://orcid.org/0000-0001-9220-2467>

Xun Xu

BGI Research <https://orcid.org/0000-0002-5338-5173>

Jian Wang

Beijing Genomics Institute

Chao Nie

BGI Research

Ming Zhao

Chinese Academy of Medical Sciences and Peking Union Medical College

Qianjin Lu

Hunan Key Laboratory of Medical Epigenomics, Second Xiangya Hospital, Central South University

Liang Xiao

BGI-Shenzhen <https://orcid.org/0000-0003-0836-4397>

Ruijin Guo

BGI-Shenzhen <https://orcid.org/0000-0001-7863-3797>

Article

Keywords:

Posted Date: October 16th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9933411/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: There is **NO** Competing Interest.

Abstract

The human skin microbiome plays a critical role in host physiology and immunity, yet its genomic and functional diversity remains under-characterized. Here, we present the Global Skin Metagenomic Data (GSMD) resource, an integrated catalog of 12,409 newly sequenced and 4,396 publicly available skin metagenomes, along with 3,077 isolate genomes. From this dataset, we reconstructed 452,397 metagenome-assembled genomes (MAGs) and identified 11,891 prokaryotic, 165 fungal, and 87,770 viral species-level genome bins (SGBs). This resource introduces 42,758 previously uncharacterized taxa, substantially expanding the known phylogenetic diversity and metabolic repertoire of skin-associated microbial communities. In addition, we cataloged 1.71 billion protein sequences, identifying 79,434 novel sequence clusters lacking homology to existing functional databases. By generating high-confidence three-dimensional structural models for 1,018 of these uncharacterized proteins, we identified unique structural folds that suggest novel biochemical functions. The GSMD resource establishes a high-resolution genomic and structural foundation that facilitates the systematic exploration and standardized interpretation of the human skin microbiome.

Introduction

The human skin is a dynamic, stratified interface colonized by a multi-kingdom microbial community of bacteria, archaea, fungi, and viruses that collectively modulate barrier integrity, immune education, and host physiology¹⁻⁴. Over the past 15 years, high-throughput sequencing has reshaped our understanding of the spatial organization, temporal stability, and individuality of skin microbial communities^{3,5,6}. Nevertheless, high-resolution mechanistic studies of the skin microbiome remain constrained by substantial gaps in reference genomes and functional annotations⁷. Compared to the human gut⁸⁻¹⁰, the skin microbiome remains under-characterized, largely due to its low microbial biomass, high host-DNA background, and the historical scarcity of diverse reference genomes².

To facilitate metagenomic analysis, recent efforts have established several skin-specific catalogs, such as the Skin Microbial Genome Collection (SMGC)⁷, the United Human Skin Genome (UHSG)¹¹ and the Early-Life Skin Genomes (ELSG)¹². While these resources have substantially improved read-mapping rates, they remain limited in capturing non-bacterial domains, global strain-level variation, and niche-specific adaptations^{7,11,12}. The skin is a nutrient-poor, acidic, and highly selective environment^{1,3,13}. Consequently, determining how commensals maintain fitness through specific functional pathways—such as amino acid, cofactor, and glycan metabolism—is critical^{11,13}. Furthermore, the skin acts as a primary interface for environmental exposure and pathogen transmission, making the comprehensive profiling of cutaneous antibiotic resistance genes (ARGs) an urgent clinical priority⁵.

A major bottleneck in current microbiome research is the vast fraction of metagenomic sequence space that constitutes "functional dark matter"—microbial proteins lacking homology to existing databases^{14,15}. Traditional sequence-alignment methods often fail to characterize these ubiquitous gene

products, leaving a massive functional repertoire unexplored¹⁶⁻¹⁸. Recent advancements in deep-learning-based structure prediction provide a powerful alternative to overcome this limitation¹⁹. By predicting three-dimensional folds directly from primary sequences, it is now possible to infer putative biochemical activities and prioritize candidates for experimental validation, bridging the gap between sequence discovery and functional validation^{17,18,20}.

To address these challenges, we present the Global Skin Metagenomic Data (GSMD) resource, an integrated multi-kingdom catalog of the human skin microbiome. By systematically processing 12,409 newly sequenced and 4,396 publicly available skin metagenomes, along with 3,077 isolate genomes, we reconstructed 452,397 metagenome-assembled genomes (MAGs). This compendium significantly expands the known phylogenetic diversity of skin microbiota, introducing 42,758 previously uncharacterized taxa across prokaryotic, fungal, and viral domains. We demonstrate that this expanded diversity broadens our understanding of critical metabolic pathways and ARG reservoirs within the skin niche. Furthermore, by cataloging 1.71 billion protein sequences, we identified 79,434 novel sequence clusters. We applied deep-learning-based structural modeling and generated high-confidence three-dimensional structures for 1,018 representative proteins. Together, GSMD provides a standardized, structurally-informed genomic resource for accelerating functional discovery and guiding future translational studies in skin microbiome research.

Results

Construction of the Global Skin Metagenomic Data (GSMD)

To establish the Global Skin Metagenomic Data (GSMD) resource, we compiled 233.33 terabytes of shotgun metagenomic data, including 12,409 newly sequenced samples from the Chinese Skin Microbiome Project (CSMP) and SZ-4D cohorts and 4,396 publicly available skin metagenomes (**Supplementary Table 1**). This global compendium spans diverse skin microenvironments, physiological states, dermatological conditions, and human populations (**Fig. 1a**). Using a standardized de novo assembly and binning pipeline, we recovered 253,565 prokaryotic metagenome-assembled genomes (pMAGs), including 124,941 high-quality and 128,624 medium-quality (**Fig. 1b**). Co-clustering these pMAGs with public reference isolates at a 95% average nucleotide identity (ANI) threshold defined 11,891 prokaryotic species-level genome bins (pSGBs) (**Supplementary Tables 2 and 3**). Extending domain-specific workflows to non-prokaryotic assemblies further recovered 6,355 fungal MAGs (fMAGs), grouped into 165 fungal species-level genome bins (fSGBs) (**Supplementary Table 4**), and 192,477 viral MAGs (vMAGs), grouped into 87,770 viral species-level genome bins (vSGBs) (**Supplementary Table 5**).

Beyond taxonomic expansion, we established a comprehensive functional catalog by predicting 1.71 billion non-redundant protein sequences across all assembled contigs, which were consolidated into 1,272,886 robust protein families (**Supplementary Table 6**). Of these, 79,434 families lacked sequence homology to existing databases (Pfam or RefSeq), representing novel functional repertoire. By enabling high-throughput 3D structural modeling of these uncharacterized families, this functional catalog

provides an unprecedented resource to uncover novel protein folds and putative catalytic sites, fundamentally expanding the structural repertoire of skin-derived bioactive molecules.

High-resolution phylogenetic and genomic expansion of the skin prokaryome

The 253,565 pMAGs recovered in GSMD span 34 phyla and cluster into 11,891 pSGBs (**Fig. 2a and Supplementary Table 3**). Four major phyla—Pseudomonadota, Actinomycetota, Bacillota, and Bacteroidota—drove the majority of this diversity, accounting for 79.22% of the pSGB richness and 92.01% of all pMAGs (**Fig. 2b and Supplementary Fig. 1**). Among them, Pseudomonadota emerged as the most diverse lineage, encompassing 88,416 pMAGs distributed across 5,102 pSGBs and 146 families (**Supplementary Fig. 1a**). At the family level, the dominant contributors within this phylum included *Sphingomonadaceae* (729 SGBs; n = 11,385 pMAGs), *Rhodobacteraceae* (285 SGBs; n = 11,149 pMAGs), *Moraxellaceae* (453 SGBs; n = 10,223 pMAGs), and *Burkholderiaceae* (525 SGBs; n = 10,098 pMAGs) (**Supplementary Fig. 1a and Supplementary Table 3**).

For canonical skin commensals, Actinomycetota accounted for 34.5% of the reconstructed genomes, largely driven by *Propionibacteriaceae* (107 SGBs; n = 35,791 pMAGs). This family includes *Cutibacterium* (19 pSGBs; n = 34,785), the most abundant skin genus^{3,5,13} and the one yielding the highest number of recovered genomes in our dataset (**Supplementary Fig. 1b**). Additionally, the genus *Corynebacterium* (from 262 pSGBs, n = 14,377) from the *Mycobacteriaceae* family (from 436 pSGBs, n = 27,676) was the second most reconstructed genus and also a relatively abundant bacterium in the skin^{3,5,13}. The most studied genus in the skin⁴, *Staphylococcus* (from 64 pSGBs, n = 12,389), was the third most reconstructed genus and came from Bacillota (from 1,183 pSGBs, n = 28,292) (**Supplementary Fig. 1c**). The dataset also highlights a notable representation of Bacteroidota within the skin microbiome. Nearly 10% of the assembled genomes (25,611 pMAGs; 1,608 pSGBs) belong to Bacteroidota. Though traditionally considered a hallmark of the gut microbiome, families such as *Weeksellaceae*, *Bacteroidaceae*, and *Porphyromonadaceae* show a broad distribution across the cutaneous landscape (**Supplementary Fig. 1d**).

By mapping the reconstructed pSGBs against the SPIRE¹⁰ and GTDB^{21,22} databases, we resolved 4,767 pSGBs (40.1% of the total 11,891) as known species (kpSGBs) containing at least one reference genome (**Fig. 2a-c and Supplementary Table 3**). The genomic recovery within these kpSGBs (202,138 pMAGs) closely mirrored the physiological distribution of our sampling cohorts: 45.38% of the genomes (n = 91,738) originated from sebaceous sites, followed by dry sites (39.23%; n = 79,305), moist sites (11.99%; n = 24,234), and sites with unspecified microenvironmental information (3.39%, n = 6,861) (**Supplementary Fig. 2a-d**). At the population level, kpSGBs were recovered across all available country-level cohorts, with the largest contribution from China (n = 183,437), followed by Australia (n = 6,554), the USA (n = 1,414), Malaysia (n = 152), Singapore (n = 98), Italy (n = 35), and India (n = 16), with an additional 276 known pMAGs from samples without country information (**Supplementary Fig. 2e**). Unsurprisingly, this known fraction of the catalog was dominated by canonical skin commensals. *Cutibacterium acnes* was the most ubiquitous species across all microenvironments, yielding 24,986

pMAGs (**Fig. 3a, c, e**). Notably, *Lawsonella clevelandensis* represented another major component of the catalog, with 10,426 recovered pMAGs, exceeding the genomic recovery of several well-recognized skin colonizers (**Fig. 3a, c, e**). This was accompanied by deep genomic recoveries of other highly prevalent colonizers, including *Cutibacterium granulosum* (n = 6,227), *Staphylococcus epidermidis* (n = 5,816), *Staphylococcus capitis* (n = 3,194), and *Corynebacterium kefirresidentii* (n = 2,323) (**Fig. 3a, c, e and Supplementary Fig. 2f**). The profound sequencing depth achieved for these kpSGBs is consistent with their known ecological prominence and provides a robust foundation for downstream strain-level analyses.

Additionally, at the archaeal level, a total of 429 archaeal pMAGs were reconstructed, spanning four phyla (Thermoproteota, Halobacteriota, Asgardarchaeota, and Methanobacteriota) and clustered into 47 pSGBs. Among these, only three pSGBs could be assigned to known species, namely *Natronococcus amylolyticus*, *Pyrobaculum aerophilum*, and *Methanobrevibacter_A sp900769095*. Another 30 pSGBs could only be annotated at the order level, mainly including JANJXX01 (from 26 SGBs, n = 334), DRAE01 (from one pSGB, n = 16), JAGGXT01 (from one pSGB, n = 1) and Wukongarchaeia (from 2 pSGBs, n = 2), whereas the remaining pSGBs could only be annotated at the family or genus level (**Fig. 2d and Supplementary Table 3**). Together, the integration of these elusive archaeal clades with deeply resolved bacterial commensals substantially expands the known phylogenetic boundaries of skin-resident prokaryotes.

Widespread uncharacterized microbial diversity across healthy and diseased skin

Beyond well-characterized taxa, a striking 59.91% of the recovered pSGBs (7,124 of 11,891) lacked reference genomes in the SPIRE or GTDB databases, delineating an extensive catalog of uncharacterized pSGBs (upSGBs; **Fig. 2a-c**). The integration of these upSGBs profoundly expanded the known phylogenetic diversity of the skin microbiome, increasing the total phylogenetic tree branch length by 76.9% (**Fig. 2a, b**). These upSGBs encompassed 51,427 pMAGs (20.28% of the total), with 13.78% remaining unclassifiable even at the genus level (**Supplementary Fig. 2a-d and Supplementary Table 3**).

Ecologically, previously uncharacterized species were widespread across skin microenvironments, populating all seven dry sites, alongside multiple moist and sebaceous sites (**Supplementary Fig. 2c, d**). To evaluate whether these upSGBs were broadly shared or specific to particular populations and ecological contexts, we quantified their unique and shared distributions across countries, skin microenvironments, and disease states. A substantial fraction was exclusive to the Chinese cohorts, with relatively few upSGBs shared across different countries. This pattern, partly reflecting the large contribution of Chinese cohorts, suggests that GSMD expands the representation of population-associated skin microbial diversity (**Supplementary Fig.3a and Supplementary Table 7**). Across skin microenvironments, 745 upSGBs were shared by dry, moist, and sebaceous sites, whereas 2,458, 536, and 1,800 pSGBs were specific to dry, moist, and sebaceous sites, respectively, indicating that the unexplored skin microbiome contains both broadly distributed and niche-specific taxa (**Supplementary Fig.3b and Supplementary Table 8**).

Despite extensive prior metagenomic profiling of sebaceous sites, 18.13% of pMAGs recovered from these niches still lacked reference genomes. This proportion was even higher in dry environments (23.4%), while moist sites harbored 17.24% uncharacterized pMAGs (**Supplementary Fig. 2a, b**). At the population level, uncharacterized pMAGs were unevenly distributed, with most species- and genus-level unknowns recovered from Chinese samples, followed by Australia, and only limited numbers from the USA and Malaysia (**Supplementary Fig. 2e**). Beyond population and physiological niches, the substantial presence of upSGBs was also evident across pathological states. We assembled uncharacterized genomes from all 10 surveyed skin diseases, where they constituted an average of 18.2% of the recovered pMAGs. Furthermore, the identification of both disease-specific and shared upSGBs suggests that a portion of this unexplored diversity may be associated with distinct disease contexts (**Supplementary Fig. 3c and Supplementary Table 9**). Notably, this uncharacterized fraction was highly enriched in samples from discoid lupus erythematosus (25.72%), vitiligo (23.06%), alopecia areata (21.46%), and androgenic alopecia (20.92%) (**Supplementary Fig. 4a, b**).

Taxonomically, Pseudomonadota dominated the unexplored fraction, contributing 26,706 pMAGs across 3,277 upSGBs (**Supplementary Fig. 4c, d**). Strikingly, our assemblies within this phylum captured several branching monophyletic clades completely absent from extant reference databases. Clades assigned to the family Micavibrionaceae, as well as uncharacterized lineages such as JAEXMB01, UBA2020, and JAIXQD01, form distinct ANI-based clusters devoid of any previously described genera or species (**Supplementary Fig. 1a and Supplementary Fig. 5**). Other prevalent upSGBs primarily mapped to the family Sphingomonadaceae (e.g., *Novosphingopyxis*) and Xanthomonadaceae (e.g., *Lysobacter*) (**Fig. 3b, d, f and Supplementary Fig. 6a**). Bacteroidota represented the second-largest reservoir of novel diversity (n = 6,793 pMAGs; 865 upSGBs), predominantly encompassing the orders Chitinophagales, Flavobacteriales, and Cytophagales (**Supplementary Fig. 6b**). Even within the well-studied Actinomycetota (n = 3,741 pMAGs; 654 upSGBs), the skin-dominant genus *Corynebacterium* alone yielded 74 uncharacterized SGBs, exposing substantial hidden diversity (**Supplementary Fig. 6c**). Crucially, the single most prevalent upSGB across the entire dataset—pSGB05262, a member of Deinococcota—was recovered in 1,259 MAGs (**Fig. 3 b, d, f, Supplementary Fig. 2e and Supplementary Fig. 6a**). Collectively, the ubiquitous distribution of these uncharted taxa across diverse physiological and pathological states substantially broadens our understanding of skin microbial ecology, highlighting a massive, untapped genetic reservoir tightly linked to cutaneous health.

Uncharacterized taxa broaden the metabolic and resistance repertoire of the skin prokaryome

To evaluate the functional potential of these unexplored taxa, we constructed a non-redundant catalog of 36,818,660 gene sequences across the 11,891 pSGBs. Functional annotation revealed a broad repertoire enriched in core metabolic pathways, including amino acid, carbohydrate, nucleotide, and energy metabolism, as well as cofactor/vitamin biosynthesis (**Supplementary Fig. 7**). Principal coordinates analysis (PCoA) of functional gene frequencies revealed a distinct functional divergence between upSGBs and kpSGBs across 12 phyla, including the skin-dominant Pseudomonadota, Actinomycetota, Bacillota, and Bacteroidota (**Fig. 4a and Supplementary Table 10**). Specifically, upSGBs within

Actinomycetota displayed a significantly higher prevalence of genes involved in amino acid, energy, cofactor/vitamin, and lipid metabolism. Similarly, upSGBs across Pseudomonadota and Bacteroidota were enriched in cofactor/vitamin and glycan metabolism, whereas uncharacterized Bacillota lineages expanded the glycan-processing repertoire, collectively broadening the functional landscape of the skin microbiome (**Fig. 4b**).

Given that the human skin microbiome constitutes a critical reservoir for antibiotic resistance genes (ARGs) that can be mobilized by opportunistic pathogens⁵, we profiled the resistome of pSGBs. We identified 22,304 ARGs distributed across 9,430 pSGBs, with predicted resistance to 153 antibiotics (**Supplementary Table 11**). Strikingly, ARGs conferring resistance to ceftolozane, bleomycin, kasugamycin, and plazomicin were exclusively encoded by upSGBs (**Supplementary Fig. 8**), exposing previously overlooked bacterial communities as significant hidden reservoirs of resistance. Taxonomically, the resistome was predominantly driven by the highly diverse Pseudomonadota, followed by Bacillota, Bacteroidota, and Actinomycetota (**Fig. 4c**). Unexplored lineages further extended this ARG diversity into historically under-sampled phyla, including Verrucomicrobiota, Patescibacteria, Cyanobacteriota, and Thermoproteota (**Fig. 4c**). These deeply embedded traits underscore how unexplored microbial lineages fundamentally shape the architecture of the skin resistome.

GSMD resolves a vast, uncharacterized skin virome and its phage-host interactome

Despite their crucial ecological roles, human skin viruses remain systematically underrepresented in reference databases^{2,23}. By mining 700 million contigs across the GSMD, we resolved an extensive skin viral catalog comprising 192,477 vMAGs (**Fig. 5a, Supplementary Table 5**). The vast majority of these genomes were bacteriophages (80.16%), primarily assigned to the Caudoviricetes (75.79%), alongside a minor fraction of proviruses (2.85%; **Supplementary Fig. 9a and Supplementary Table 5**). More than half of the vMAGs (54.08%; genome sizes up to 380.6 kb) were classified as high-quality or complete, underscoring the robustness of our viral genome recovery (**Supplementary Fig. 9b, c**). This viral repertoire, largely driven by the 169,195 vMAGs from our newly sequenced metagenomes, spanned diverse skin microenvironments, including sebaceous (76,868 vMAGs), dry (75,994 vMAGs) and moist (33,986 vMAGs) niches. Furthermore, these viral taxa were widely detected across 10 dermatological conditions, ranging from vitiligo (31,704 vMAGs) to subacute cutaneous lupus erythematosus (318 vMAGs) (**Fig. 5b, c, Supplementary Fig. 9d**).

To delineate viral taxonomy, we clustered the vMAGs following MIUViG standards (95% ANI, 85% alignment fraction)²⁴, resolving them into 87,770 vSGBs (**Fig. 5d and Supplementary Table 5**). Cross-referencing against comprehensive public databases revealed that 35,606 vSGBs (40.5%) were entirely novel to this study (**Fig. 5e**). Ecological comparisons underscored a pronounced host-specific adaptation: the skin virome exhibited greater overlap with other human-associated viromes (gut^{25,26}, oral²⁷, and reproductive tract²⁸) than with environmental sources. Despite this shared host signature, the skin virome maintained strong niche specificity, with 49,914 vSGBs remained entirely absent from non-skin human body sites. Furthermore, 62,570 were not detected from massive environmental databases,

including NCBI RefSeq, Global Ocean Virome 2.0 (GOV 2.0)²⁹, and freshwater and soil viromes from the Integrated Microbial Genomes (IMG) database³⁰ (**Supplementary Fig. 9e**). Functionally, the skin virome represented a large pool of uncharacterized genetic dark matter; only 9.49% of the vSGBs harbored proteins annotatable via the EggNOG viral database, predominantly encoding enzymatic, metabolic, and DNA replication functions (**Supplementary Fig. 9f**).

Mapping predicted interactions between skin phages and bacterial hosts revealed a dense phage-host network of 605,238 predicted linkages, connecting 59.28% of the skin vSGBs to 61.15% of the reconstructed pSGBs (**Supplementary Fig. 10a**). We further examined whether uncharacterized bacterial species were also incorporated into this phage-host network. At the linkage level, 127,660 predicted phage-host connections involved upSGBs, whereas 477,578 involved kpSGBs (**Supplementary Fig. 10b**). At the host level, phage connectivity was significantly lower among upSGBs than among kpSGBs (Fisher's exact test, $p < 0.001$); specifically, 3,451 of 7,124 upSGBs (48.44%) and 3,821 of 4,767 kpSGBs (80.16%) were linked to at least one phage (**Supplementary Table 12**). Nevertheless, nearly half of the upSGBs showed evidence of phage linkage, indicating that previously unexplored bacterial lineages are also subject to phage-mediated ecological regulation.

These interactions were predominantly specific, with 86.29% restricted to the genus level (**Fig. 5f**). While phage connectivity was broadly distributed across major skin phyla (55.4% - 72.9% prevalence in Pseudomonadota, Bacteroidota, Actinomycetota, and Bacillota), it was notably enriched within Bacillota (**Supplementary Fig. 10c**). Specifically, 79 of the top 100 pSGBs most heavily targeted by phages belonged to Bacillota, dominated by the clinically relevant genera *Staphylococcus* (26 pSGBs) and *Streptococcus* (53 pSGBs) (**Supplementary Fig. 10d**). The remaining 21 highly targeted pSGBs mapped entirely to *Moraxella* within Pseudomonadota. Collectively, these dynamics highlight a largely unexplored skin virome where phages actively exert selective pressure on key commensal and pathogenic bacterial populations.

Genomic resolution of the skin mycobiome reveals *Malassezia* predominance

Profiling fungal constituents of the skin microbiome has historically been challenging due to the low microbial biomass and high host-DNA background characteristic of skin samples^{31,32}. Deep sequencing of the GSMD cohort enabled assembly-based recovery of fMAG. We reconstructed 6,355 medium- to high-quality fMAGs, delineating a non-redundant catalog of 165 fSGBs (**Fig. 6a and Supplementary Table 4**). These recovered genomes provide a high-resolution genomic baseline, delineating the lineage diversity that populates the cutaneous environment. The composition of recovered fSGBs confirmed that the genus *Malassezia* is the dominant genus of the skin mycobiome (**Fig. 6a and Supplementary Table 4**). Notably, *M. restricta* and *M. globosa* exhibited the highest frequency of genomic recovery, with their fMAGs constituting the majority of the reconstructed fungal fraction across diverse skin microenvironments (**Fig. 6b**). This broad genomic prevalence highlights their ubiquity across different human populations. Furthermore, our assembly-based approach successfully captured 28 uncharacterized fSGBs that currently lack representatives in isolate-based databases. These

uncharacterized fSGBs were distributed across seven fungal classes, including Eurotiomycetes (n = 12), Malasseziomycetes (n = 6), Pucciniomycetes (n = 2), Dipodascomycetes (n = 3), Pichiomycetes (n = 2), Dothideomycetes (n = 2), and Exobasidiomycetes (n = 1), underscoring the extensive, previously hidden taxonomic diversity of the skin mycobiome (**Supplementary Fig. 11**).

To characterize the functional repertoire of the recovered skin mycobiome, we annotated 1,754,318 protein-coding genes derived from the 165 fSGBs. Mapping against the eggNOG³³ and KEGG³⁴ databases assigned putative functions to 39.25% and 27.71% of this catalog, respectively (**Supplementary Table 13**). PCoA based on orthologous group frequencies revealed a distinct functional divergence that closely mirrored the phylogenetic architecture of the fungal taxa, distinctly segregating the Ustilaginomycotina, Pezizomycotina, and Saccharomycotina subphyla (**Fig. 6c**). To further dissect these subphylum-level differences, we assessed KEGG module completeness across 45 high-quality fSGBs to map their core metabolic pathways (**Supplementary Fig. 12 and Supplementary Table 4**). While essential pathways governing amino acid, lipid, carbohydrate, glycan, energy, and nucleotide metabolism were universally conserved across all species, we observed clade-specific metabolic specializations. Notably, pathways driving the biosynthesis of complex secondary metabolites, such as penicillin, paspaline, chanoclavine aldehyde, and fumigaclavine, were exclusively enriched within Pezizomycotina (**Supplementary Fig. 12**). This distinct biosynthetic capacity suggests that Pezizomycotina lineages may possess a specialized arsenal of secondary metabolites, potentially mediating unique ecological interactions or adaptation strategies within the skin microenvironment.

The uncharacterized functional landscape of the skin microbiome

To systematically chart the functional landscape of the cutaneous ecosystem, we predicted over 1.71 billion protein sequences across 16,805 metagenomic datasets. By collapsing this massive sequence space using established stringent clustering thresholds¹⁴, we consolidated 200.2 million non-redundant proteins into 1,272,886 robust sequence clusters (≥ 100 members per family) suitable for high-confidence structural modeling (**Fig. 1b, Supplementary Table 6**). Strikingly, by systematically filtering against existing reference databases (Pfam and RefSeq), we identified 79,434 completely unannotated sequence clusters. These clusters, hereafter referred to as skin novel metagenomic protein families (SNMPFs), represent a massive, untapped functional reservoir within the skin microbiome (**Fig. 1b**).

We next sought to link SNMPFs to their phylogenetic origins by mapping their source contigs against our multi-kingdom genomic catalogs. Of the 14.6 million contigs encoding SNMPF members, a substantial fraction remained entirely unclassifiable (59.3%; n = 8,680,173), further highlighting the vast unmapped genetic space in this ecosystem. Among the taxonomically assigned contigs, the vast majority were of bacterial origin (n = 5,013,775), alongside notable contributions from fungi (n = 860,316), viruses (n = 46,058), and archaea (n = 37,646). Notably, the composition of these protein families frequently transcended taxonomic boundaries. The largest subset comprised mixed bacterial and taxonomically unassigned sequences, followed by complex bacterial-fungal-unassigned admixtures, whereas strictly archaeal or viral clusters were rare (**Fig. 7a**).

Rather than representing transient or rare genetic elements, these novel protein families are deeply entrenched within the skin microbiome architecture. The vast majority of SNMPFs (89.69%) were broadly shared across two or more SGBs, with 70.8% exhibiting wide distribution across distinct phyla (**Fig. 7b**). Conversely, a smaller subset (18.6%) demonstrated strict species- or genus-level specificity, suggesting niche-specialized adaptation. Ecologically, these SNMPFs were highly enriched in canonical skin colonizers, spanning dominant bacterial phyla (Pseudomonadota, Actinomycetota, Bacillota, and Bacteroidota) and fungal divisions (Basidiomycota and Ascomycota) (**Supplementary Fig. 13a**). At higher taxonomic resolution, we pinpointed profound SNMPF enrichment within keystone skin genera—such as the fungus *Malassezia* and bacteria *Cutibacterium*, *Staphylococcus*, and *Corynebacterium* (**Supplementary Fig. 13b**). Collectively, these findings underscore that despite exhaustive genomic reconstruction, an immense proportion of the functional repertoire actively deployed by the core skin microbiota remains uncharacterized.

Deep structural modeling unveils novel enzymes in the skin microbiome

To bridge the gap between sequence novelty and biological function, we leveraged ESMFold²⁰ to resolve the three-dimensional structures of all 79,434 SNMPFs. This high-throughput structural profiling yielded 1,018 high-confidence protein models (predicted pLDDT > 90; **Supplementary Table 14**). By interrogating these topologies for binding cavities and utilizing DeepFRI for structure-based functional annotation, we detected 6,709 potential pockets. Applying stringent quality filters (score > 60) refined this pool to 85 high-quality active sites across 808 functionally annotated proteins (**Supplementary Tables 15 and 16**). Functional mapping of this subset revealed a striking predominance of transport-related activities (**Fig. 7c**), suggesting these novel proteins frequently mediate critical molecular exchanges or nutrient acquisition within the cutaneous ecosystem.

To achieve near-atomic resolution for these highest-confidence candidates, we subjected the 74 proteins harboring the 85 high-quality pockets to rigorous structural re-evaluation using AlphaFold3¹⁷. This advanced modeling confirmed 30 robust structural topologies (pLDDT > 80, pTM > 0.8, ranking_score > 0.85) featuring 38 refined, high-confidence cavities (**Supplementary Tables 17 and 18**). Integrating these precise structural coordinates with deep homology analyses allowed us to assign putative enzymatic roles to these previously uncharacterized proteins.

Notably, this high-resolution mapping revealed two distinct families of previously unrecognized monooxygenases. The first adopts a canonical P450-like architecture, complete with a central α -helical bundle and a highly conserved heme-binding pocket, consistent with its predicted catalytic function. Conversely, the second monooxygenase family preserves the core enzymatic scaffold but incorporates two spatially distinct high-confidence pockets, pointing toward an evolutionary adaptation for multi-substrate recognition or complex allosteric regulation (**Fig. 7d, e**). Parallel analyses pinpointed two putative one-carbon transferases characterized by classic α/β folds, where crucial functional residues are intricately clustered around highly confident pocket structures (**Fig. 7f, g**).

Together, these structural insights translate unmapped sequence space into biochemically tractable targets, establishing a concrete foundation for future skin-derived enzymatic innovation and microbiome-targeted interventions.

Discussion

The human skin microbiome represents a highly complex and dynamic ecosystem, yet its exploration at the genomic and functional levels has historically lagged behind other major body sites^{7,8,35}. In this study, we introduced the GSMD resource, a comprehensive compendium that integrates over 16,000 metagenomes to dramatically expand the known phylogenetic and functional diversity of the cutaneous microbiome. By recovering over 452,000 MAGs across multiple kingdoms and delineating more than 42,000 previously uncharacterized species-level genome bins (SGBs), GSMD substantially alleviates the reference-genome bottleneck in skin microbiome research. This expansion not only enhances read-mapping rates for future metagenomic surveys but also unveils a vast, unexplored microbial landscape associated with human skin across diverse microenvironments and health states.

An important outcome of our analysis is the profound uncharacterized diversity embedded within the skin microbiome. Uncharacterized pSGBs were not restricted to rare taxa or isolated niches; instead, they were broadly distributed across sebaceous, dry, and moist microenvironments and were consistently detected in both healthy and diseased skin. These lineages substantially expanded the phylogenetic breadth of the cutaneous prokaryome and contributed distinct metabolic capacities, including pathways related to amino acid, lipid, cofactor, glycan, and energy metabolism. From a clinical perspective, detection of ARGs within uncharacterized taxa further indicates that current reference-dependent approaches may underestimate the complexity and clinical relevance of the skin resistome^{5,36}. These findings support a revised view in which unexplored microbial lineages represent recurrent and functionally relevant components of the skin ecosystem.

GSMD also improves genomic resolution of non-bacterial components of the skin microbiome. The recovered viral catalog expanded current knowledge of the skin virome, revealing substantial viral diversity that remains underrepresented in existing human-associated and environmental viral databases^{37,38}. Host-prediction analyses further uncovered a dense phage-host interaction network, particularly involving Bacillota genera such as *Staphylococcus* and *Streptococcus*, suggesting that phages may contribute to ecological regulation and genome diversification of both commensal and opportunistic pathogenic populations^{39,40}. In parallel, the fungal genome catalog confirmed the predominance of *Malassezia* while identifying uncharacterized fungal SGBs and clade-specific metabolic specializations^{2,41}. These findings emphasize that the skin microbiome should be understood as a multi-kingdom ecosystem shaped by bacterial, archaeal, fungal, and viral interactions^{2,4,5,13}.

Beyond genome reconstruction, GSMD provides a structure-guided strategy for interrogating the functional dark matter of the skin microbiome^{14,42}. By clustering 1.71 billion predicted proteins, we identified tens of thousands of novel protein families lacking detectable similarity to Pfam or RefSeq.

Structural modeling, pocket prediction, and structure-based functional inference enabled prioritization of high-confidence candidates with putative functional sites, including previously unrecognized monooxygenase-like and one-carbon transferase-like proteins^{17-19,43}. This sequence-to-structure-to-function framework moves microbiome analysis beyond cataloging unknown genes toward generating experimentally actionable hypotheses for microbial metabolism, host interaction, and skin-derived enzyme discovery.

Despite these advances, we acknowledge several limitations inherent to this work. First, while MAGs provide essential blueprints of uncultivated taxa, they remain computational inferences that can suffer from fragmentation or mis-binning, particularly concerning mobile genetic elements and accessory genomes. Although we enforced rigorous quality control, the complete genomic architecture and physiological behavior of these novel taxa ultimately necessitate experimental isolation. Second, our functional annotations and small-molecule binding predictions serve as *in silico* hypotheses. Translating these putative interactions into validated biological mechanisms will require extensive biochemical characterization and *in vivo* models. Furthermore, although our sampling spans diverse geographic and clinical conditions, the sheer heterogeneity of the global human population implies that the true boundaries of cutaneous microbial diversity remain to be fully charted.

In conclusion, the GSMD resource establishes a robust, multi-kingdom foundation for the next generation of skin microbiome research. By illuminating the uncharacterized phylogenetic diversity, structural protein landscape, and complex trans-kingdom interactomes of the cutaneous ecosystem, we provide a vital catalyst for moving from descriptive metagenomics toward mechanistic understanding and targeted therapeutic intervention. We anticipate that GSMD will serve as an enduring community resource, driving future discoveries in microbial ecology, personalized dermatology, and precision medicine.

Methods

Collection of new skin metagenomes

We included two newly generated datasets, CSMP and SZ-4D, both of which were sequenced using the DNBSEQ-T1 platform and are being made publicly available concurrently with this study. The CSMP dataset comprised 9,958 skin metagenomic samples collected from volunteers aged 10–65 years who were recruited from the Hospital of Dermatology, Chinese Academy of Medical Sciences; the Second Xiangya Hospital of Central South University; the Second People's Hospital of Chengdu; Hainan Hospital of Dermatology; the Second Affiliated Hospital of Harbin Medical University; and the People's Hospital of Tibet Autonomous Region. These samples covered multiple geographic regions of China, including central, eastern, southwestern, northeastern, and southern areas, and encompassed 10 skin diseases, 11 skin sites, and three skin microenvironments: sebaceous, dry, and moist. The SZ-4D dataset included 2,451 skin metagenomic samples collected from healthy individuals recruited in Shenzhen, Guangdong Province, covering four skin sites across two skin microenvironments: sebaceous and moist. Detailed

information on these newly generated datasets is provided in Table S1. The study protocol was approved by the Ethics Review Committee of the Hospital of Dermatology, Chinese Academy of Medical Sciences, and the Institutional Review Board of BGI. Written informed consent was obtained from all participants.

Publicly available skin metagenomic datasets

We collected publicly available skin metagenomic samples from 9 different studies^{3,5,11-13,44-48}, comprising 4,396 metagenomes and 68.56Tb of high-quality skin metagenomic data. These samples represented seven countries, including the United States (1099 samples), Australia (631 samples), China (2393 samples), and additional countries (273 samples). The metagenomes were obtained from 20 skin sites across three skin environments: sebaceous, dry, and moist. Despite manual curation, metadata describing the sampling site and skin microenvironment remained unavailable for 852 samples in the public repositories included in this study. All manually curated metadata fields, including sample origin, geographic location, sampling site, skin microenvironment, disease status, and study accession, are provided in Supplementary Table 1. In addition, we retrieved 3,077 genomes from three culture-based studies⁴⁹⁻⁵¹ for inclusion in the constructed genome collection (**Supplementary Table 2**).

DNA extraction and metagenomic sequencing

Sample collection for each volunteer was performed by trained dermatologists. On the day of sampling, volunteers were instructed not to use cleansing or moisturizing products. For each sampling site, a sterile swab pre-moistened with sterile saline was gently rubbed back and forth 50 times over an 8 cm² area for approximately 30 s under standardized conditions. The swabs were then placed in DNA preservation solution and promptly stored at -80°C until DNA extraction.

DNA was extracted using the MagPure Stool DNA KF Kit B (MD5115, Magen), and DNA concentration was quantified using a Qubit fluorometer (Invitrogen). Given the typically low biomass of skin samples, sequencing libraries were prepared using the MGIEasy Fast FS DNA Library Prep Set (MGI, Cat. No. 940-000030-00) on the BGI NGS platform. Library construction was performed with 1–10 ng of extracted DNA as input, following the manufacturer's recommended protocol. The resulting libraries were sequenced on the DNBSEQ-T1 platform to generate paired-end 150-bp reads.

Preprocessing of metagenomic data

Quality control of raw metagenomic reads was performed using fastp⁵² (v0.20.1) with stringent filtering and trimming criteria, retaining reads with an average Phred quality score ≥ 20 and a length ≥ 30 bp. Host-derived reads were subsequently removed by aligning the quality-filtered reads against the human reference genome GRCh38 using Bowtie2⁵³ (v2.4.2) with the '--end-to-end --very-sensitive' options. Reads mapped to the human genome were discarded, and the remaining non-human reads were retained for downstream analyses. As a result, approximately 14.22 Gb reads of high-quality non-human metagenomic data were obtained for further analysis (**Supplementary Table 1**).

Metagenomic assembly and extraction of MAGs

De novo metagenomic assembly was performed independently for each metagenome to recover MAGs. Clean non-human reads from each sample were assembled using MEGAHIT⁵⁴ (v1.2.9) with default parameters. Contigs shorter than 1,000 bp were discarded, resulting in approximately 700 million contigs with a total length of 1.75 trillion nucleotides. Sequencing reads were then mapped back to the assembled contigs using Bowtie2⁵³ (v2.4.2) to calculate contig coverage profiles. Genome binning was performed using MetaBAT2⁵⁵, MaxBin2⁵⁶, and CONCOCT⁵⁷ implemented in the MetaWRAP⁵⁸ (v1.3) suite. The resulting bins were further refined and optimized using DAS Tool⁵⁹ (v1.1.7).

Quality control, dereplication, and species-level clustering of prokaryotic genomes

Genome quality of prokaryotic MAGs was assessed using CheckM2⁶⁰ (v1.0.2) with default parameters. Chimerism and taxonomic contamination were evaluated using GUNC⁶¹ (v1.0.6) with the 'gunc_db_progenomes2.1dmnd' database. Genomes were excluded if they had completeness <50%, contamination ≥10%, a quality score, defined as completeness - 5×contamination, <50, or a GUNC clade separation score ≥0.45. In total, 253,565 pMAGs passed the quality-control criteria and were retained for downstream analysis. Among them, 124,941 high-quality pMAGs were defined as those with completeness ≥90% and contamination <5%, and 128,624 medium-quality pMAGs were defined as those with completeness ≥50% and contamination <10%. Dereplication of the 253,565 pMAGs was performed using dRep⁶² (v3.4.4) with the parameters “-pa 0.9 -sa 0.95 -nc 0.3 --S_algorithm fastANI”, resulting in 11,891 pSGBs.

Taxonomic classification and phylogenetic analysis of skin prokaryotic genomes

Taxonomic annotation of pSGBs was performed using the Genome Taxonomy Database Toolkit^{21,22} (GTDB-Tk, v2.0.4) with the GTDB r220 database. Protein sequence alignments generated by GTDB-Tk were used for phylogenetic reconstruction. Maximum-likelihood phylogenetic trees were inferred using IQ-TREE2⁶³ (v2.3.5) for 11,844 bacterial and 47 archaeal species. “ModelFinder” was used to automatically select the best-fit substitution model according to the Bayesian Information Criterion (BIC). The LG+F+R10 model was used for bacterial phylogeny construction, whereas the LG+F+R4 model was selected for archaeal phylogeny reconstruction. The resulting phylogenetic trees were visualized using the R package ggtree⁶⁴.

Comparison of pSGBs with Representative Species in the SPIRE Database

To assess the novelty of the skin surface genomes relative to existing prokaryotic genomes, all constructed pSGBs were compared with 133,402 reference prokaryotic species from the Searchable, Planetary-scale Microbiome Resource (SPIRE)¹⁰ database. Pairwise genome comparisons between the two genome collections were performed using fastANI⁶⁵ with the parameter ‘--minFraction 0.3’, as implemented in dRep⁶², to calculate genome coverage and average nucleotide identity (ANI). A pSGB

was considered represented in SPIRE if it matched a SPIRE reference genome with genome coverage $\geq 30\%$ and ANI $\geq 95\%$. pSGBs that did not match any SPIRE reference genome and were not assigned to a known species by GTDB-Tk^{21,22} were considered upSGBs.

Functional annotation and identification of antibiotic resistance genes in pSGBs

Coding sequences (CDSs) of all pSGBs were predicted using Prodigal⁶⁶ (v2.6.3) with parameters ‘-p single, -f gff, -a, -d’. A total of 36,818,660 CDSs were identified across 11,891 pSGBs. Functional annotation was performed using EggNOG-mapper³³ (v2.12) based on EggNOG⁶⁷ data (v5.0.2). Antibiotic resistance genes (ARGs) were identified using Resistance Gene Identifier (RGI)³⁶ (v6.0.3) based on the Comprehensive Antibiotic Resistance Database (CARD)³⁶ database (v3.2.9). In total, 22,304 ARGs were annotated across 9,430 pSGBs.

Identification and quality control of viral genomes

Viral genome identification was performed using assembled contigs longer than 5,000 bp. Putative viral contigs were first identified using the end-to-end module of geNomad⁶⁸ (v1.7.6), retaining only contigs with geNomad scores > 0.7 . Genome quality was then assessed using the end-to-end workflow of CheckV⁶⁹ (v1.0.3). Viral genomes classified as medium-quality, high-quality, or complete were retained whereas viral genomes with $> 30\%$ host-gene content were removed. To further assess potential cellular contamination, hmmsearch was used with default parameters to identify benchmarking universal single-copy orthologs (BUSCOs)⁷⁰ from bacteria, archaea, and fungi within the putative viral genomes. The BUSCO ratio was defined as the number of BUSCOs divided by the total number of genes in each viral genome, and genomes with bacterial, archaeal, or fungal BUSCO ratio $\geq 5\%$ were excluded from further analysis. After quality control, a total of 192,477 vMAGs were obtained, including 62,462 complete, 41,624 high-quality (HQ), and 88,391 medium-quality (MQ) vMAGs.

Genome clustering and phylogenetic tree construction

The vMAGs were clustered into 87,770 vSGBs using BLASTn⁷¹ v2.12.0 with parameters ‘-evalue 1e-10 -word_size 20’, based on thresholds of 85% alignment coverage and 95% nucleotide identity. Genomes within each cluster were ranked using the following scoring formula: completeness – contamination + log (genome size), and the genome with the highest score was selected as the representative genome for that cluster. The same alignment parameters were used to assess sequence overlap between our dataset and various viral genome databases, including infant-gut (0–7 years)²⁶, oral virome²⁷, vaginal virome²⁸, gut virome²⁵, soil virome³⁰, freshwater virome³⁰, marine virome²⁹, and RefSeq viral genome (version 2024-10-5) databases. To investigate phylogenetic relationships among viral genomes, a viral proteome tree was constructed using ViPTreeGen⁷² (v1.1.2) with default parameters. Treemmer⁷³ (v0.3) was then used to assess redundancy in the phylogenetic tree and reduce it to 15,998 branches.

Taxonomic classification of the vSGBs

Taxonomic assignment of viral genomes was performed based on geNomad⁶⁸ classification results, which follows the ICTV taxonomy as defined in the Master Species List 39 (MSL39). Viral genomes were classified to the family level where possible, but no genus- or species-level assignments were made within families. In addition, annotations were manually curated to remove taxonomically implausible assignments.

Host prediction for the vSGBs

Host prediction for vSGBs was performed using iPHoP⁷⁴ (version 1.3.3) with default parameters. iPHoP integrates multiple approaches in a two-step framework to predict genus-level hosts for viruses infecting bacteria and archaea while controlling the false discovery rate. A custom host database was constructed from the 11,891 pSGBs, following the procedure described in the iPHoP documentation under 'Adding Bacterial and/or Archaeal MAGs to the Host Database.'

Gene prediction and functional annotation of vSGBs

CDSs of representative skin vSGBs were predicted using Prodigal⁶⁶ v2.11.0-gv and subsequently translated into protein sequences. In total, 4,863,817 genes were predicted from the 87,770 vSGBs. Functional annotation was performed using EggNOG-mapper³³ (v2.12) against the viral database from EggNOG⁶⁷ (v5.0.2).

Identification of fungal MAGs

Fungal genomes were identified from assembled bins that were not classified as bacterial. Genome quality was assessed using Eukcc⁷⁵ (v2.1.3), and bins with completeness <50% or contamination ≥10% were removed. In total, 6,355 fMAGs passed the quality threshold and were retained for downstream analysis. Among them, 3,417 medium-quality fMAGs were defined as those with completeness ≥50% and contamination <10%, while 2,938 high-quality fMAGs were defined as those with completeness ≥90% and contamination <5%.

Taxonomic classification and clustering of fMAGs

The fMAGs were dereplicated using dRep⁶² v3.4.0 with the parameters 'dereplicate -pa 0.9 -sa 0.95 -nc 0.3 --S_algorithm fastANI'. Following previous studies on fungal genomics⁷⁶, fungal genomes from the skin surface were clustered into species-level clusters based on a 95% ANI threshold, resulting in 165 fSGBs. Taxonomic classification was performed using the NCBI fungal genome database (v. 2024-10-5). All constructed fSGBs were compared against 22,125 reference fungal species in the database using pairwise genome comparisons with fastANI⁶⁵ (--minFraction 0.3) as implemented in dRep. Based on the ANI results, fSGBs were assigned to known species when ANI to a reference genome was >95%, and to known genera when ANI was >80%.

Phylogenetic analysis and functional annotation of fSGBs

Protein-coding genes in the fSGBs were predicted using GeneMark-ES⁷⁷ (v4.68) with the parameters ‘--fungus --ES --min_contig 5000’, resulting in 1,756,358 predicted proteins. A phylogenetic tree of the fSGBs was constructed using OrthoFinder⁷⁸ (v3.0.1b1) with default parameters based on the predicted protein sequences from representative fSGB genomes. OrthoFinder identifies single-copy orthologous genes and uses them to infer phylogenetic relationships. Functional annotation was performed using eggNOG-mapper (v2.12) against the fungal database from eggNOG (v5.0.2).

Skin microbial gene catalog

Putative protein-coding sequences were predicted from all contigs using Prodigal⁶⁶ v2.6.3 with the parameters ‘-p single, -f gff, -a, and -d’, resulting in 1,713,137,422 predicted sequences. Following previously reported methods¹⁴, all-versus-all pairwise alignments were performed using LAST alignment⁷⁹. Sequence clustering was conducted using the HipMCL algorithm⁸⁰ with an inflation parameter of 2.0. A 90% sequence similarity threshold and an 80% alignment coverage threshold were applied during clustering to ensure high-quality alignments and reduce potential artifacts, including severely truncated or incomplete genes, pseudogenes, and chimeric sequences. In addition, only MCL clusters containing at least 100 members were retained to further improve the quality of the final dataset and minimize the influence of potential spurious gene products. Ultimately, 1,272,886 sequence clusters were obtained.

Identification of novel protein families

For the 1,272,886 protein sequences clusters, consensus sequences were searched against the Pfam-A database⁸¹ (v.37.0) using HMMER⁸² (v.3.1, thresholds: 7.0 total, 5.0 domain). In parallel, sequences were aligned against RefSeq (version 2024-10-5) using DIAMOND⁸³ (30% identity, 50% coverage bidirectionally). These searches identified that 539,080 clusters with significant hits in the Pfam-A database, and 1,175,442 clusters matched 1,509,627 RefSeq entries. In total, 79,434 clusters had not hits in either the Pfam-A and RefSeq databases and were therefore considered skin novel metagenome protein families (SNMPF). Based on the inferred taxonomic origins of contigs from prokaryotes, fungi, and viruses in the skin metagenomes, protein sequences were taxonomically annotated.

Structure, function, and pocket prediction of SNMPFs

Protein structures for all SNMPFs were predicted using the ESM Metagenomic Atlas²⁰. The mean pLDDT was calculated for each predicted structure, and structures with mean pLDDT > 90 were retained for downstream analysis. In total, 1,018 SNMPF structures met this criterion (**Supplementary Table 14**).

Binding pockets were predicted using AutoSite⁴³, and pockets with an empirical composite AutoSite score >60 were selected for further analysis. Putative functions and catalytic residues were predicted using DeepFRI¹⁹, which was used to infer enriched GO terms, EC numbers, and residue-level significance weights across GO categories, including biological process (BP), cellular component (CC), and molecular

function (MF). Predicted pockets and functional sites were visually inspected using PyMOL⁸⁴. In total, 74 SNMPF protein structures were retained for subsequent analysis.

For the 74 SNMPF protein structures, structures were re-predicted using AlphaFold3 Server¹⁷ and filtered using more stringent confidence criteria: pLDDT > 80, pTM > 0.8, and ranking_score > 0.85. A total of 30 protein structures passed these filtering criteria. Pocket detection and functional prediction were then repeated as described above, and PyMOL was used to visualize the predicted pockets and functional sites.

Declarations

Data and code availability

Clean metagenomic reads newly generated from the CSMP cohort have been deposited in the CNGBdb under accession number CNP0005495. The SZ-4D dataset is available in CNGBdb under accession numbers CNP0004678, CNP0004766, CNP0004207, and CNP0004677. The GSMD-derived assembled prokaryotic and fungal genomes have been deposited in CNGBdb under accession numbers CNP0006963 and CNP0009082, respectively. The assembled viral genomes are currently available at https://drive.google.com/file/d/1LsS3COj_J_r5AbOqXaYhvKkv2fPv5K21/view?usp=sharing and will be deposited in a permanent public repository before publication. Accession numbers and data access information for all publicly available datasets used in this study are provided in Supplementary Table S1. The workflows used for genome reconstruction, genome and protein catalog construction, and functional annotation are available at <https://github.com/lizhiming11/GSMD>.

Acknowledgments

This work was supported by the Noncommunicable Chronic Diseases-National Science and Technology Major Project (2026ZB0556800). National Key R&D Program of China (No. 2023YFC3603300, 2022YFC3601800), Science and Technology Innovation 2030 Major Projects 2022ZD0211600, the CAMS Innovation Fund for Medical Sciences (2019-I2M-5-066), and the Shanghai Municipal Science and Technology Major Project (2023SHZDZX02), the Major Program of the National Natural Science Foundation of China (82595960), the Key Program of National Natural Science Foundation of China (NO. 82430102), the CAMS Innovation Fund for Medical Sciences (CIFMS) No.2021-I2M-1-059, the Non-profit Central Research Institute Fund of Chinese Academy of Medical Sciences (2020-RC320-003). We also thank the assistance of State Key Laboratory of Genome and Multi-omics Technologies, Shenzhen Key Laboratory of Neurogenomics.

Author contributions

Z.M. and R.G. conceived the study. Z.M. performed multi-domain genome assembly, clustering, annotation, and catalog construction; compared the study catalog with other genome collections; constructed the protein catalog; compared it with Pfam-A and RefSeq; predicted protein structures,

binding sites, and potential functions; and drafted and revised the manuscript. Y.J. contributed to analytical discussions, assembled the SZ-4D data, performed data deposition and public release, and revised the manuscript. B.Z., Y.Z., Q.L., M.Z., and Q.L. provided the CSMP data and phenotype information. S.L., T.Q., and R.G. contributed to analytical discussions. S.L. and R.G. contributed to manuscript drafting and final revision. H.Z., S.J., X.J., and J.W. provided valuable comments on the draft. S.Z., T.Z., X.J., W.Z., X.X., J.W., and N.C. provided feedback and funding support. C.N., M.Z., Q.L., L.X., and R.G. supervised the study and provided feedback and funding support. All authors read, edited and approved the final manuscript.

Declaration of interests

The authors declare no competing interests.

References

1. Grice, E.A., and Segre, J.A. (2011). The skin microbiome. *Nature reviews microbiology* 9, 244-253.
2. Byrd, A.L., Belkaid, Y., and Segre, J.A. (2018). The human skin microbiome. *Nature Reviews Microbiology* 16, 143-155.
3. Oh, J., Byrd, A.L., Park, M., Kong, H.H., and Segre, J.A. (2016). Temporal stability of the human skin microbiome. *Cell* 165, 854-866.
4. Oh, J., and Voigt, A.Y. (2025). The human skin microbiome: from metagenomes to therapeutics. *Nature Reviews Microbiology*, 1-17.
5. Li, Z., Xia, J., Jiang, L., Tan, Y., An, Y., Zhu, X., Ruan, J., Chen, Z., Zhen, H., and Ma, Y. (2021). Characterization of the human skin resistome and identification of two microbiota cutotypes. *Microbiome* 9, 1-18.
6. Grice, E.A., Kong, H.H., Conlan, S., Deming, C.B., Davis, J., Young, A.C., Program, N.C.S., Bouffard, G.G., Blakesley, R.W., and Murray, P.R. (2009). Topographical and temporal diversity of the human skin microbiome. *science* 324, 1190-1192.
7. Saheb Kashaf, S., Proctor, D.M., Deming, C., Saary, P., Hölzer, M., Taylor, M.E., Kong, H.H., Segre, J.A., Almeida, A., and Finn, R.D. (2022). Integrating cultivation and metagenomics for a multi-kingdom view of skin microbiome diversity and functions. *Nature microbiology* 7, 169-179.
8. Pasolli, E., Asnicar, F., Manara, S., Zolfo, M., Karcher, N., Armanini, F., Beghini, F., Manghi, P., Tett, A., and Ghensi, P. (2019). Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell* 176, 649-662. e620.
9. Consortium, T.H.M.P. (2012). Structure, function and diversity of the healthy human microbiome. *nature* 486, 207-214.
10. Schmidt, T.S., Fullam, A., Ferretti, P., Orakov, A., Maistrenko, O.M., Ruscheweyh, H.-J., Letunic, I., Duan, Y., Van Rossum, T., and Sunagawa, S. (2024). SPIRE: a Searchable, Planetary-scale microbiome REsource. *Nucleic Acids Research* 52, D777-D783.

11. Li, Z., Ju, Y., Xia, J., Zhang, Z., Zhen, H., Tong, X., Sun, Y., Lu, H., Zong, Y., and Chen, P. (2023). Integrated Human Skin Bacteria Genome Catalog Reveals Extensive Unexplored Habitat-Specific Microbiome Diversity and Function. *Advanced Science* 10, 2300050.
12. Shen, Z., Robert, L., Stolpman, M., Che, Y., Allen, K.J., Saffery, R., Walsh, A., Young, A., Eckert, J., and Deming, C. (2023). A genome catalog of the early-life human skin microbiome. *Genome biology* 24, 252.
13. Oh, J., Byrd, A.L., Deming, C., Conlan, S., Kong, H.H., and Segre, J.A. (2014). Biogeography and individuality shape function in the human skin metagenome. *Nature* 514, 59-64.
14. Pavlopoulos, G.A., Baltoumas, F.A., Liu, S., Selvitopi, O., Camargo, A.P., Nayfach, S., Azad, A., Roux, S., Call, L., and Ivanova, N.N. (2023). Unraveling the functional dark matter through global metagenomics. *Nature* 622, 594-602.
15. Baltoumas, F.A., Karatzas, E., Liu, S., Ovchinnikov, S., Sofianatos, Y., Chen, I.M., Kyrpides, N.C., and Pavlopoulos, G.A. (2024). NMPFamsDB: a database of novel protein families from microbial metagenomes and metatranscriptomes. *Nucleic Acids Res* 52, D502-d512. 10.1093/nar/gkad800.
16. Radivojac, P., Clark, W.T., Oron, T.R., Schnoes, A.M., Wittkop, T., Sokolov, A., Graim, K., Funk, C., Verspoor, K., Ben-Hur, A., et al. (2013). A large-scale evaluation of computational protein function prediction. *Nat Methods* 10, 221-227. 10.1038/nmeth.2340.
17. Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A.J., and Bambrick, J. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* 630, 493-500.
18. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., and Potapenko, A. (2021). Highly accurate protein structure prediction with AlphaFold. *nature* 596, 583-589.
19. Gligorijević, V., Renfrew, P.D., Kosciolk, T., Leman, J.K., Berenberg, D., Vatanen, T., Chandler, C., Taylor, B.C., Fisk, I.M., and Vlamakis, H. (2021). Structure-based protein function prediction using graph convolutional networks. *Nature communications* 12, 3168.
20. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., and Shmueli, Y. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 379, 1123-1130.
21. Chaumeil, P.-A., Mussig, A.J., Hugenholtz, P., and Parks, D.H. (2020). GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics*.
22. Chaumeil, P.-A., Mussig, A.J., Hugenholtz, P., and Parks, D.H. (2022). GTDB-Tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics* 38, 5315-5316.
23. Chen, Y.E., Fischbach, M.A., and Belkaid, Y. (2018). Skin microbiota–host interactions. *Nature* 553, 427-436.
24. Roux, S., Adriaenssens, E.M., Dutilh, B.E., Koonin, E.V., Kropinski, A.M., Krupovic, M., Kuhn, J.H., Lavigne, R., Brister, J.R., and Varsani, A. (2019). Minimum information about an uncultivated virus genome (MIUViG). *Nature biotechnology* 37, 29-37.

25. Nayfach, S., Páez-Espino, D., Call, L., Low, S.J., Sberro, H., Ivanova, N.N., Proal, A.D., Fischbach, M.A., Bhatt, A.S., and Hugenholtz, P. (2021). Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. *Nature microbiology* 6, 960-970.
26. Peng, Y., Zhu, J., Wang, S., Liu, Y., Liu, X., DeLeon, O., Zhu, W., Xu, Z., Zhang, X., and Zhao, S. (2024). A metagenome-assembled genome inventory for children reveals early-life gut bacteriome and virome dynamics. *Cell Host & Microbe* 32, 2212-2230. e2218.
27. Li, S., Guo, R., Zhang, Y., Li, P., Chen, F., Wang, X., Li, J., Jie, Z., Lv, Q., and Jin, H. (2022). A catalog of 48,425 nonredundant viruses from oral metagenomes expands the horizon of the human oral virome. *Iscience* 25.
28. Huang, L., Guo, R., Li, S., Wu, X., Zhang, Y., Guo, S., Lv, Y., Xiao, Z., Kang, J., and Meng, J. (2024). A multi-kingdom collection of 33,804 reference genomes for the human vaginal microbiome. *Nature Microbiology* 9, 2185-2200.
29. Gregory, A.C., Zayed, A.A., Conceição-Neto, N., Temperton, B., Bolduc, B., Alberti, A., Ardyna, M., Arkhipova, K., Carmichael, M., and Cruaud, C. (2019). Marine DNA viral macro-and microdiversity from pole to pole. *Cell* 177, 1109-1123. e1114.
30. Camargo, A.P., Nayfach, S., Chen, I.-M.A., Palaniappan, K., Ratner, A., Chu, K., Ritter, S.J., Reddy, T., Mukherjee, S., and Schulz, F. (2023). IMG/VR v4: an expanded database of uncultivated virus genomes within a framework of extensive functional, taxonomic, and ecological metadata. *Nucleic acids research* 51, D733-D743.
31. Santiago-Rodriguez, T.M., Le François, B., Macklaim, J.M., Doukhanine, E., and Hollister, E.B. (2023). The Skin Microbiome: Current Techniques, Challenges, and Future Directions. *Microorganisms* 11. 10.3390/microorganisms11051222.
32. Jo, J.H., Kennedy, E.A., and Kong, H.H. (2017). Topographical and physiological differences of the skin mycobiome in health and disease. *Virulence* 8, 324-333. 10.1080/21505594.2016.1249093.
33. Cantalapiedra, C.P., Hernández-Plaza, A., Letunic, I., Bork, P., and Huerta-Cepas, J. (2021). eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Molecular biology and evolution* 38, 5825-5829.
34. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic acids research* 45, D353-D361.
35. Almeida, A., Nayfach, S., Boland, M., Strozzi, F., Beracochea, M., Shi, Z.J., Pollard, K.S., Sakharova, E., Parks, D.H., Hugenholtz, P., et al. (2021). A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nature Biotechnology* 39, 105-114. 10.1038/s41587-020-0603-3.
36. Alcock, B.P., Huynh, W., Chalil, R., Smith, K.W., Raphenya, A.R., Wlodarski, M.A., Edalatmand, A., Petkau, A., Syed, S.A., and Tsang, K.K. (2023). CARD 2023: expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database. *Nucleic acids research* 51, D690-D699.
37. Li, Z., Li, S., Han, C., Chen, Y., Zhen, H., Sun, Y., Zhou, X., Chen, Y., Zheng, Y., Han, L., et al. (2025). A comprehensive reference catalog of human skin DNA virome reveals novel viral diversity and

- microenvironmental influences. *Microbiol Spectr* 13, e0117825. 10.1128/spectrum.01178-25.
38. Graham, E.H., Tom, W.A., Neujahr, A.C., Adamowicz, M.S., Clarke, J.L., Herr, J.R., and Fernando, S.C. (2023). The persistence and stabilization of auxiliary genes in the human skin virome. *Viol J* 20, 49. 10.1186/s12985-023-02012-3.
 39. Alsaadi, S.E., Lu, H., Zhang, M., Dykes, G.F., Allison, H.E., and Horsburgh, M.J. (2024). Bacteriophages from human skin infecting coagulase-negative *Staphylococcus*: diversity, novelty and host resistance. *Sci Rep* 14, 8245. 10.1038/s41598-024-59065-9.
 40. Hannigan, G.D., Meisel, J.S., Tyldsley, A.S., Zheng, Q., Hodgkinson, B.P., SanMiguel, A.J., Minot, S., Bushman, F.D., and Grice, E.A. (2015). The human skin double-stranded DNA virome: topographical and temporal diversity, genetic enrichment, and dynamic associations with the host microbiome. *mBio* 6, e01578-01515. 10.1128/mBio.01578-15.
 41. Findley, K., Oh, J., Yang, J., Conlan, S., Deming, C., Meyer, J.A., Schoenfeld, D., Nomicos, E., Park, M., Kong, H.H., and Segre, J.A. (2013). Topographic diversity of fungal and bacterial communities in human skin. *Nature* 498, 367-370. 10.1038/nature12171.
 42. Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N.N., Anderson, I.J., Cheng, J.F., Darling, A., Malfatti, S., Swan, B.K., Gies, E.A., et al. (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature* 499, 431-437. 10.1038/nature12352.
 43. Ravindranath, P.A., and Sanner, M.F. (2016). AutoSite: an automated approach for pseudo-ligands prediction—from ligand-binding sites identification to predicting key ligand atoms. *Bioinformatics* 32, 3142-3149.
 44. Liu, Y., Li, Z., Zhang, C., Li, B., Jiang, W., Li, H., Zhang, L., Zhen, H., Bao, S., and Li, X. (2025). A human skin microbiome reference catalog and the skin microbial landscape of plateau adults. *iMetaOmics*, e70000.
 45. Chia, M., Naim, A.N., Tay, A.S., Lim, K., Chew, K.L., Yow, S.J., Chen, J., Common, J.E., Nagarajan, N., and Tham, E.H. (2022). Shared signatures and divergence in skin microbiomes of children with atopic dermatitis and their caregivers. *Journal of Allergy and Clinical Immunology* 150, 894-908.
 46. Tay, A.S., Li, C., Nandi, T., Chng, K.R., Andiappan, A.K., Mettu, V.S., de Cevins, C., Ravikrishnan, A., Dutertre, C.-A., and Wong, X.C.C. (2021). Atopic dermatitis microbiomes stratify into ecologic dermatotypes enabling microbial virulence and disease severity. *Journal of Allergy and Clinical Immunology* 147, 1329-1340.
 47. Chng, K.R., Tay, A.S.L., Li, C., Ng, A.H.Q., Wang, J., Suri, B.K., Matta, S.A., McGovern, N., Janela, B., and Wong, X.F.C.C. (2016). Whole metagenome profiling reveals skin microbiome-dependent susceptibility to atopic dermatitis flare. *Nature microbiology* 1, 1-10.
 48. Tett, A., Pasolli, E., Farina, S., Truong, D.T., Asnicar, F., Zolfo, M., Beghini, F., Armanini, F., Jousson, O., and De Sanctis, V. (2017). Unexplored diversity and strain-level structure of the skin microbiome associated with psoriasis. *NPJ biofilms and microbiomes* 3, 14.
 49. Yu, T., Xu, X., Liu, Y., Wang, X., Wu, S., Qiu, Z., Liu, X., Pan, X., Gu, C., and Wang, S. (2024). Multi-omics signatures reveal genomic and functional heterogeneity of *Cutibacterium acnes* in normal and

- diseased skin. *Cell Host & Microbe* 32, 1129-1146. e1128.
50. Zhou, W., Spoto, M., Hardy, R., Guan, C., Fleming, E., Larson, P.J., Brown, J.S., and Oh, J. (2020). Host-specific evolutionary and transmission dynamics shape the functional diversification of *Staphylococcus epidermidis* in human skin. *Cell* 180, 454-470. e418.
51. Ahle, C.M., Stødkilde, K., Poehlein, A., Bömeke, M., Streit, W.R., Wenck, H., Reuter, J.H., Hüpeden, J., and Brüggemann, H. (2022). Interference and co-existence of staphylococci and *Cutibacterium acnes* within the healthy human skin microbiome. *Communications biology* 5, 923.
52. Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34, i884-i890.
53. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods* 9, 357-359.
54. Li, D., Liu, C.-M., Luo, R., Sadakane, K., and Lam, T.-W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 31, 1674-1676.
55. Kang, D.D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., and Wang, Z. (2019). MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* 7, e7359.
56. Wu, Y.-W., Simmons, B.A., and Singer, S.W. (2016). MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* 32, 605-607.
57. Alneberg, J., Bjarnason, B.S., De Bruijn, I., Schirmer, M., Quick, J., Ijaz, U.Z., Lahti, L., Loman, N.J., Andersson, A.F., and Quince, C. (2014). Binning metagenomic contigs by coverage and composition. *Nature methods* 11, 1144-1146.
58. Uritskiy, G.V., DiRuggiero, J., and Taylor, J. (2018). MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* 6, 158.
59. Sieber, C.M., Probst, A.J., Sharrar, A., Thomas, B.C., Hess, M., Tringe, S.G., and Banfield, J.F. (2018). Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nature microbiology* 3, 836-843.
60. Chklovski, A., Parks, D.H., Woodcroft, B.J., and Tyson, G.W. (2023). CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nature methods* 20, 1203-1212.
61. Orakov, A., Fullam, A., Coelho, L.P., Khedkar, S., Szklarczyk, D., Mende, D.R., Schmidt, T.S., and Bork, P. (2021). GUNC: detection of chimerism and contamination in prokaryotic genomes. *Genome biology* 22, 178.
62. Olm, M.R., Brown, C.T., Brooks, B., and Banfield, J.F. (2017). dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through dereplication. *The ISME journal* 11, 2864-2868.
63. Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., Von Haeseler, A., and Lanfear, R. (2020). IQ-TREE 2: new models and efficient methods for phylogenetic inference in the

- genomic era. *Molecular biology and evolution* 37, 1530-1534.
64. Yu, G. (2020). Using ggtree to visualize data on tree-like structures. *Current protocols in bioinformatics* 69, e96.
65. Jain, C., Rodriguez-R, L.M., Phillippy, A.M., Konstantinidis, K.T., and Aluru, S. (2018). High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nature communications* 9, 5114.
66. Hyatt, D., Chen, G.-L., LoCascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics* 11, 119.
67. Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S.K., Cook, H., Mende, D.R., Letunic, I., Rattei, T., and Jensen, L.J. (2019). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic acids research* 47, D309-D314.
68. Camargo, A.P., Roux, S., Schulz, F., Babinski, M., Xu, Y., Hu, B., Chain, P.S., Nayfach, S., and Kyrpides, N.C. (2024). Identification of mobile genetic elements with geNomad. *Nature biotechnology* 42, 1303-1312.
69. Nayfach, S., Camargo, A.P., Schulz, F., Eloë-Fadrosh, E., Roux, S., and Kyrpides, N.C. (2021). CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nature biotechnology* 39, 578-585.
70. Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31, 3210-3212.
71. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. *BMC bioinformatics* 10, 421.
72. Nishimura, Y., Yoshida, T., Kuronishi, M., Uehara, H., Ogata, H., and Goto, S. (2017). ViPTree: the viral proteomic tree server. *Bioinformatics* 33, 2379-2380.
73. Menardo, F., Loiseau, C., Brites, D., Coscolla, M., Gygli, S.M., Rutaihua, L.K., Trauner, A., Beisel, C., Borrell, S., and Gagneux, S. (2018). Treemmer: a tool to reduce large phylogenetic datasets with minimal loss of diversity. *BMC bioinformatics* 19, 164.
74. Roux, S., Camargo, A.P., Coutinho, F.H., Dabdoub, S.M., Dutilh, B.E., Nayfach, S., and Tritt, A. (2023). iPhoP: An integrated machine learning framework to maximize host prediction for metagenome-derived viruses of archaea and bacteria. *PLoS biology* 21, e3002083.
75. Saary, P., Mitchell, A.L., and Finn, R.D. (2020). Estimating the quality of eukaryotic genomes recovered from metagenomic analysis with EukCC. *Genome biology* 21, 244.
76. Yan, Q., Li, S., Yan, Q., Huo, X., Wang, C., Wang, X., Sun, Y., Zhao, W., Yu, Z., and Zhang, Y. (2024). A genomic compendium of cultivated human gut fungi characterizes the gut mycobiome and its relevance to common diseases. *Cell* 187, 2969-2989. e2924.

77. Ter-Hovhannisyan, V., Lomsadze, A., Chernoff, Y.O., and Borodovsky, M. (2008). Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. *Genome research* 18, 1979-1990.
78. Emms, D.M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome biology* 20, 238.
79. Kielbasa, S.M., Wan, R., Sato, K., Horton, P., and Frith, M.C. (2011). Adaptive seeds tame genomic sequence comparison. *Genome research* 21, 487-493.
80. Azad, A., Pavlopoulos, G.A., Ouzounis, C.A., Kyrpides, N.C., and Buluç, A. (2018). HipMCL: a high-performance parallel implementation of the Markov clustering algorithm for large-scale networks. *Nucleic acids research* 46, e33-e33.
81. Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., and Mistry, J. (2014). Pfam: the protein families database. *Nucleic acids research* 42, D222-D230.
82. Eddy, S.R. (2011). Accelerated profile HMM searches. *PLoS computational biology* 7, e1002195.
83. Buchfink, B., Xie, C., and Huson, D.H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature methods* 12, 59-60.
84. DeLano, W.L. (2002). Pymol: An open-source molecular graphics tool. *CCP4 Newsl. protein crystallogr* 40, 82-92.

Figures

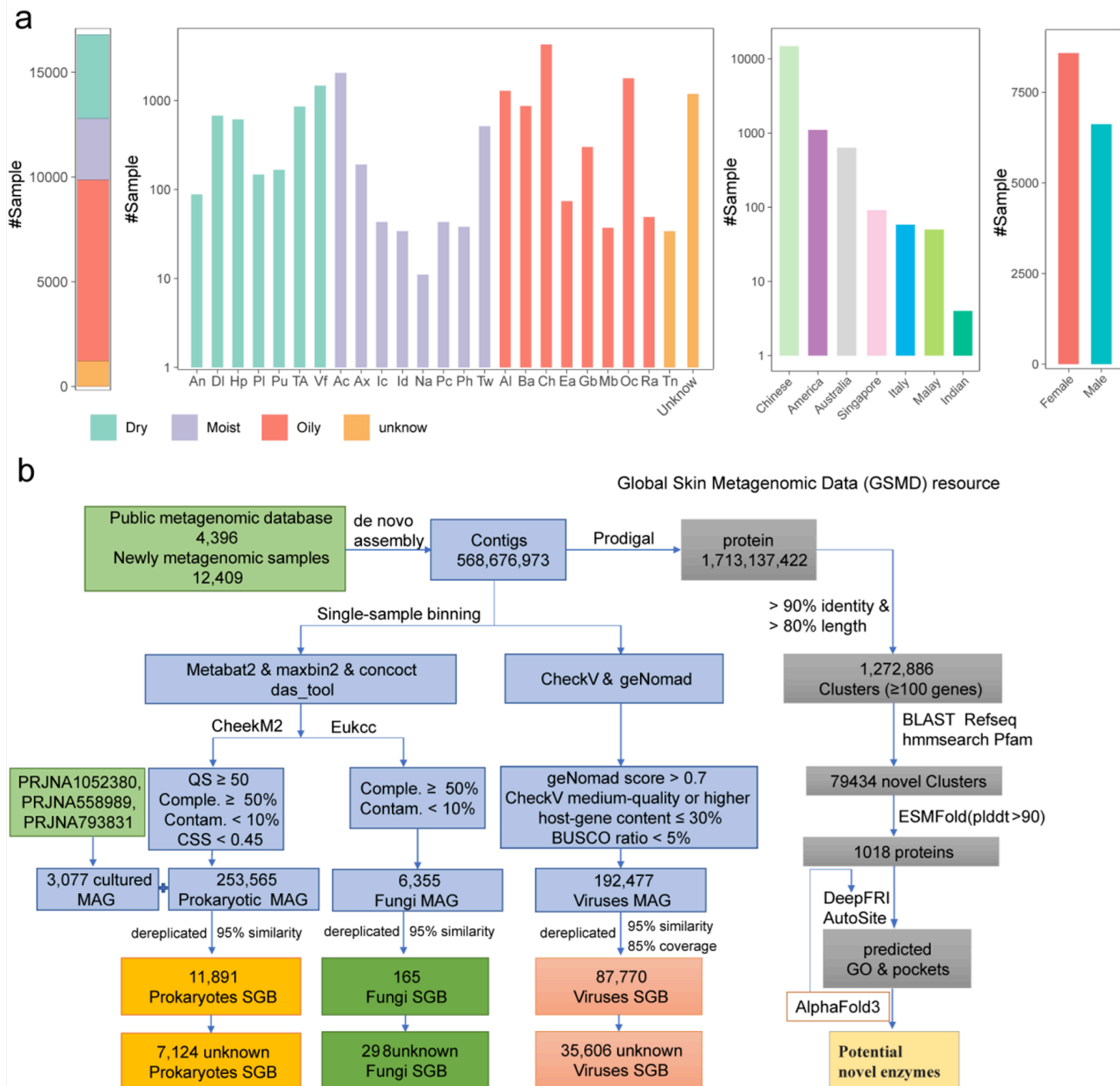


Figure 1

Construction of the GSMD resource and overview of sample. **a.** Overview of sample. Left stacked bar indicates the distribution of samples across oily, dry, moist, and “unknown” environments. Middle panel shows counts per body site within each environment category. Right panels summarize geographic representation (selected regions/countries) and sex distribution. An: area around the navel, DI: distal lower limb, Hp: hand palm, PI: proximal lower limb, Pu: proximal upper limb, TA: thoracoabdominal area, Vf: volar forearm, Ac: antecubital crease, Ax: axilla, Ic: inguinal crease, Id: interdigital digits, Na: nasal aperture, Pc: popliteal crease, Ph: plantar heel, Tw: toe web space, Al: alar line, Ba: back, Ch: cheek, Ea:

external auditory canal, Gb: glabella, Mb: midline breastbone, Oc: occipital center, Ra: retroauricular area, Tn: toenail. **b.** Overview of the GSMD pipeline: (i) integration of newly generated skin metagenomes (CSMP and SZ-4D) with public datasets and culture-derived isolate genomes; (ii) standardized preprocessing, assembly, and single-sample binning; (iii) quality control and dereplication to generate MAGs and SGBs across prokaryotes, viruses, and fungi; (iv) large-scale protein prediction and clustering, identification of SNMPFs, and structure/pocket-based functional annotation.

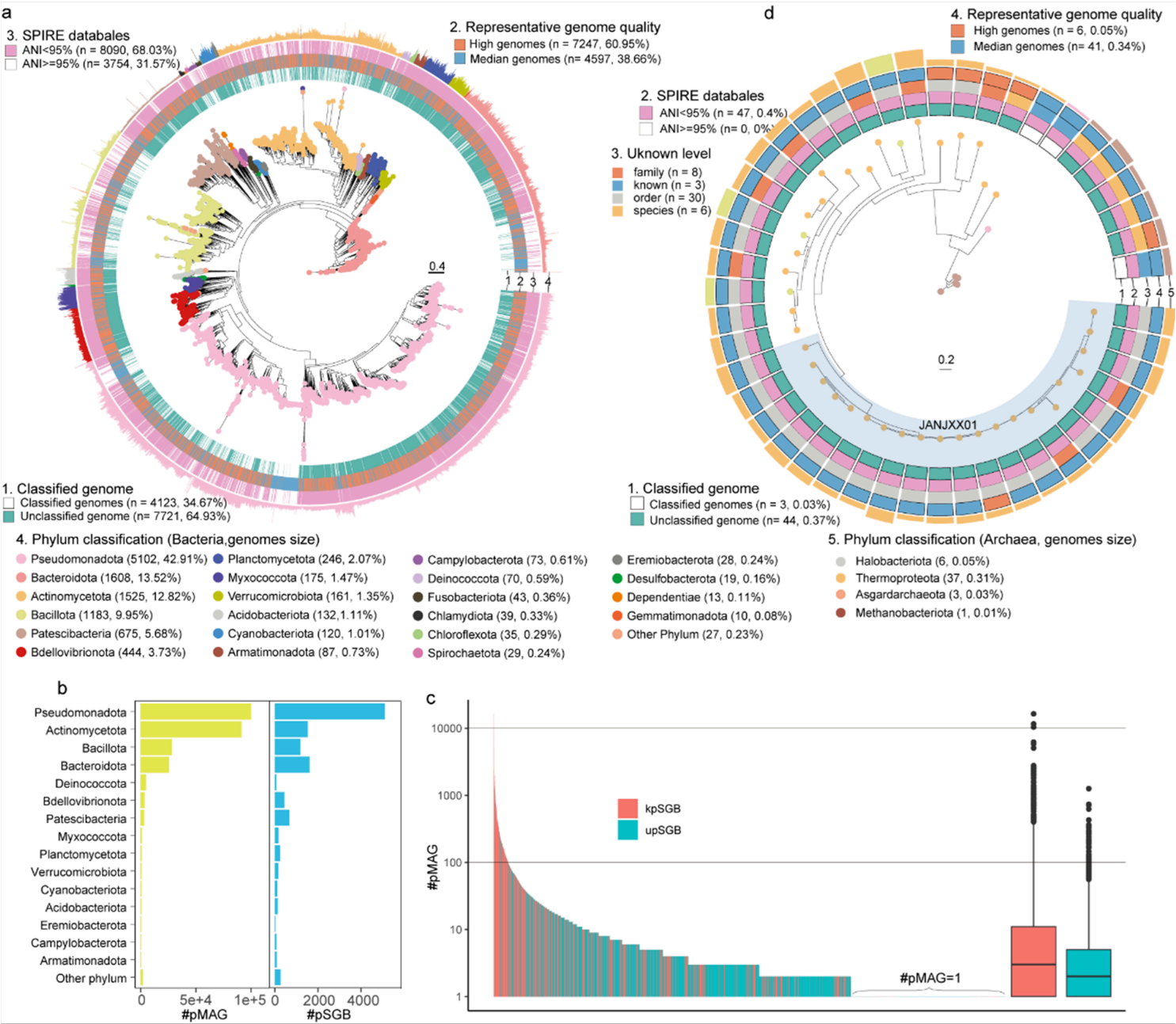


Figure 2

Phylogenetic tree of the 11,891 pSGBs detected in skin metagenomes. **a.** Phylogeny of representative genomes from bacterial SGBs reconstructed from skin metagenomes. Each leaf is a pSGB and is colored by phylum. Concentric rings (from inside to outside) indicate: classification status (classified vs unclassified to a named species), representative genome quality (High vs Medium), similarity to

SPIRE/GTDB references based on ANI ($\text{ANI} \geq 95\%$ vs $\text{ANI} < 95\%$), and genome sizes of pSGBs across different phyla. **b.** Phylum-level contributions to catalog size. Left, numbers of pMAGs per phylum; right, numbers of pSGBs per phylum. **c.** Distribution of pMAG counts by pSGB uncharacterized level. The boxplots summarize the number of MAGs per SGB for kpSGBs and upSGBs. **d.** Phylogeny of representative genomes from archaeal SGBs reconstructed from skin metagenomes. Concentric rings (from inside to outside) indicate: classification status (classified vs. unclassified to a named species), similarity to SPIRE/GTDB references based on ANI ($\text{ANI} \geq 95\%$ vs. $\text{ANI} < 95\%$), representative genome unknown level, representative genome quality (high vs. medium); and genome sizes of pSGBs across different phyla.

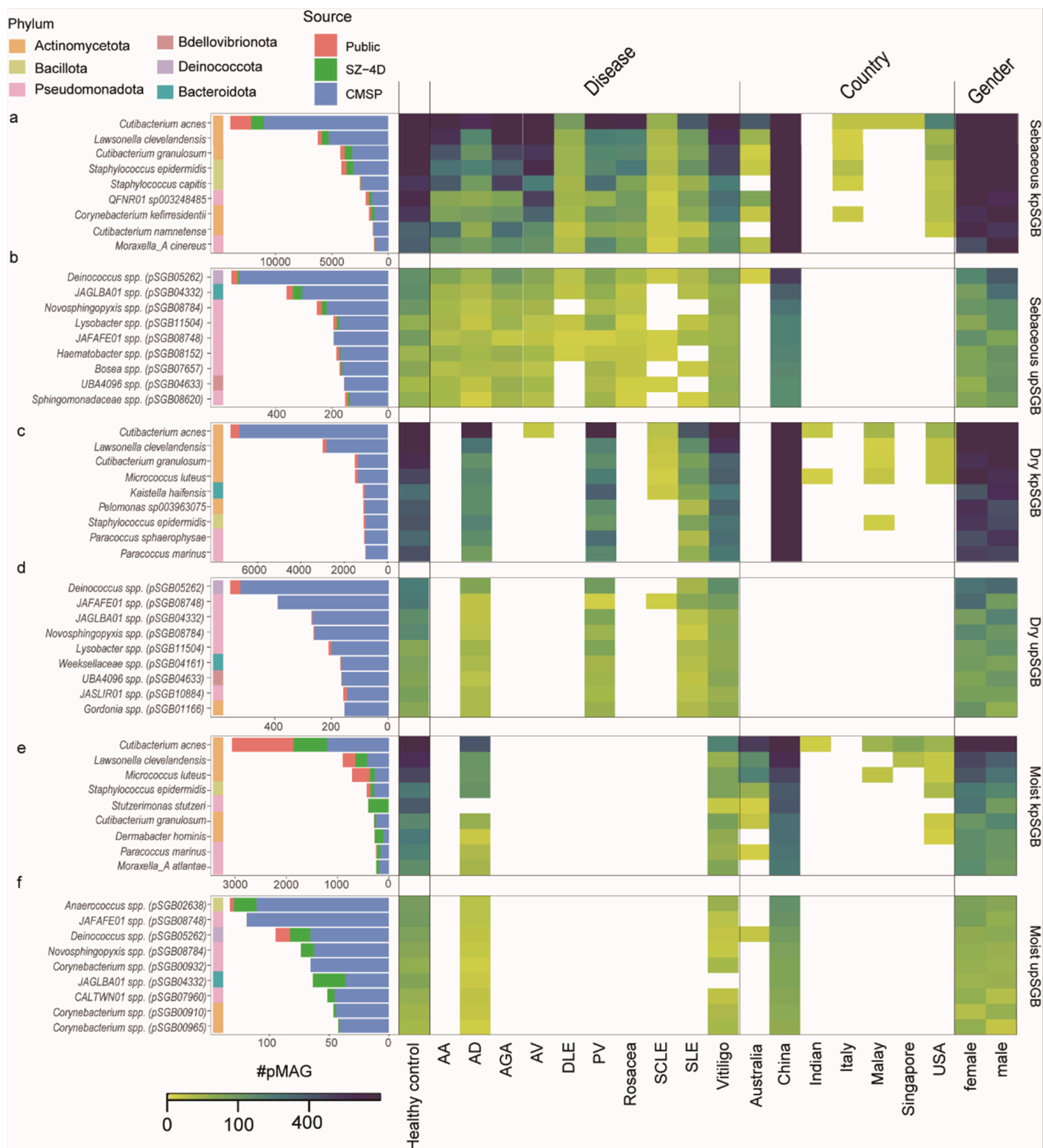


Figure 3

Distribution of pMAGs across skin environments and disease categories. The nine pSGBs with the highest numbers of pMAGs in (a, c, e) kpSGBs (pSGBs with at least one reference genome) and (b, d, f) upSGBs (pSGBs composed only of pMAGs), drawn from (a, b) sebaceous, (c, d) dry, and (e, f) moist skin sites. In each panel, pSGBs are ordered by the number of pMAGs from the corresponding skin environment. For each pSGB, the total number of skin-environment pMAGs is reported, together with

counts across 10 disease categories, 7 countries, and both genders. AA: Alopecia areata, AD: Atopic dermatitis, AGA: Androgenetic alopecia, AV: Acne vulgaris, DLE: Discoid lupus erythematosus, PV: Psoriasis vulgaris, SCLE: Subacute cutaneous lupus erythematosus, SLE: Systemic lupus erythematosus.

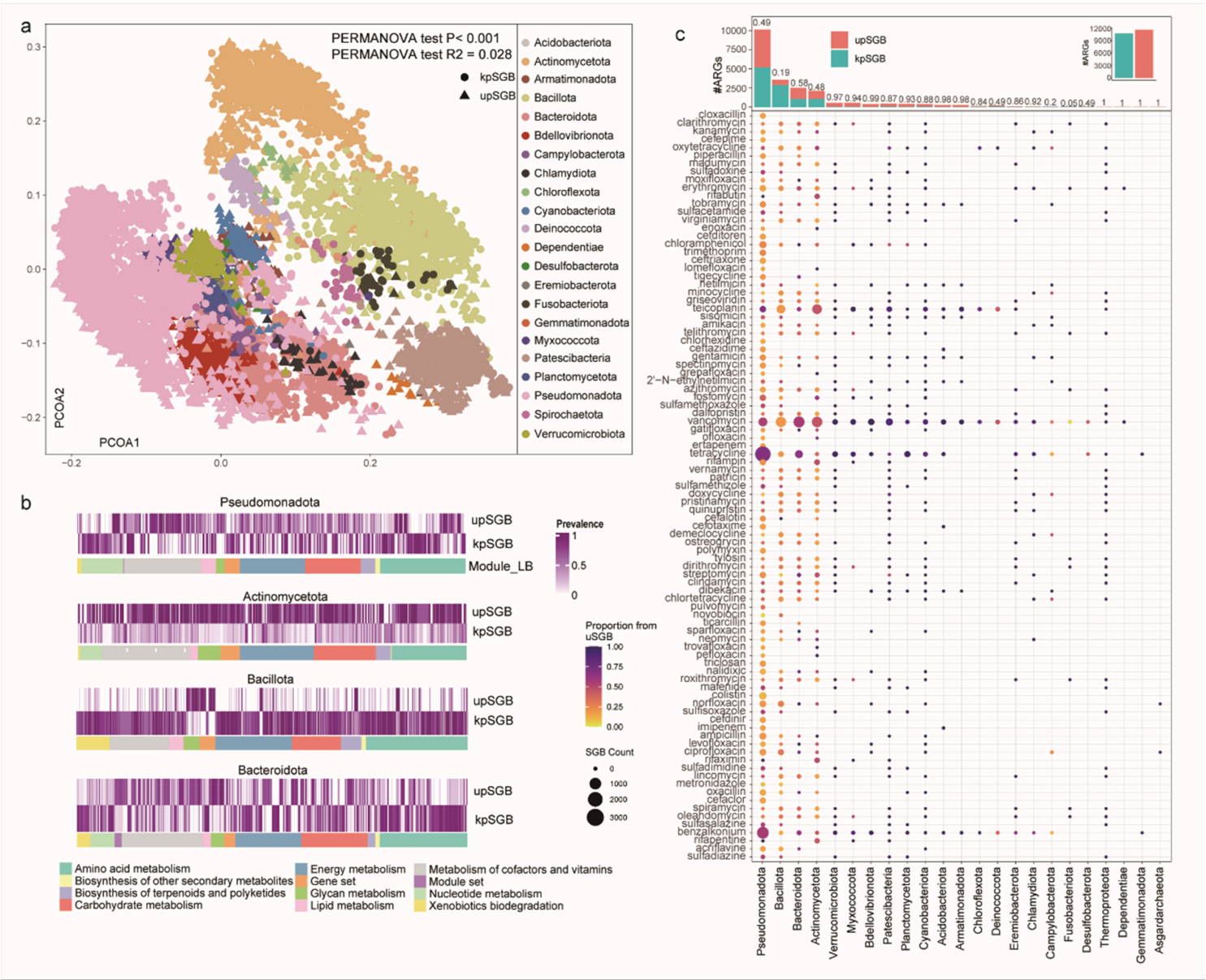


Figure 4

Differences in the functional and resistance repertoires of skin pSGBs between kpSGBs and upSGBs. a. PCoA of pSGB functional profiles based on KEGG. Each point is a pSGB colored by phylum; triangles denote upSGBs and circles denote kpSGBs. **b.** Prevalence of functions that differ between kpSGBs and upSGBs across the major phyla. The functional prevalence of pSGBs, stratified by upSGBs and kpSGBs, is shown for the four major skin phyla (Pseudomonadota, Actinomycetota, Bacillota, and Bacteroidota). The color bar at the bottom indicates different metabolic categories. **c.** Antibiotic resistance gene (ARG) landscape. The bubble plot shows the proportion of predicted ARG in upSGBs versus kpSGBs: purple

indicates a higher proportion and yellow a lower proportion. Bubble size represents the number of pSGBs harboring each predicted ARG.

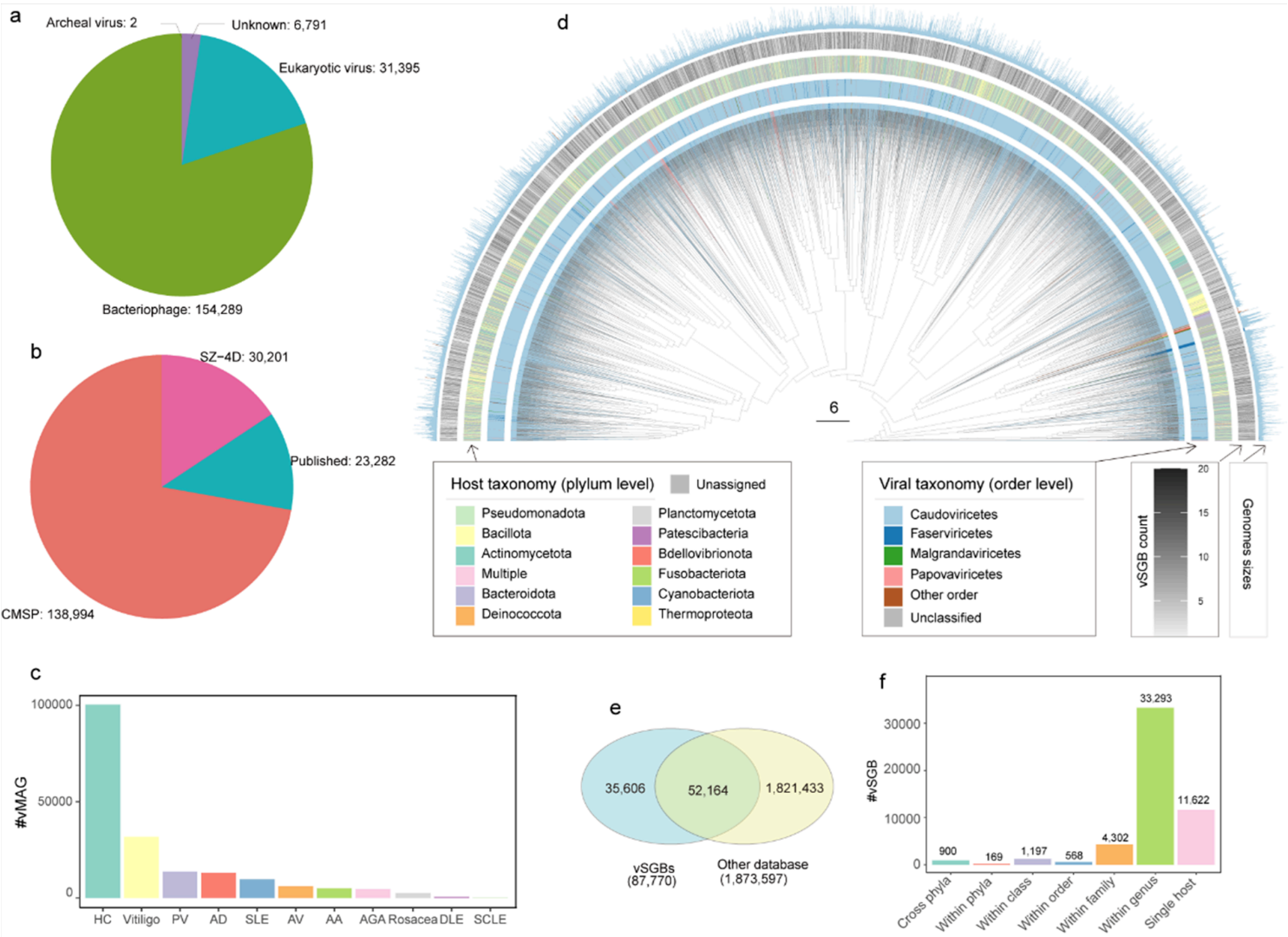


Figure 5

Diversity, sources, hosts, and novelty of skin vSGBs. **a.** Taxonomic composition of vMAGs from skin metagenomes. Pie chart shows counts assigned as bacteriophages, fungal viruses, archaeal viruses, and unknown. **b.** Most viruses come from newly sequenced samples. Pie chart partitions vMAGs recovered from the CSMP, SZ-4D, and previously published datasets. **c.** Number of vMAGs in healthy samples and across different disease categories. **d.** Phylogenomic tree of representative vSGBs. Outer rings annotate predicted bacterial host phylum (or unassigned), viral order-level taxonomy, vSGB counts per clade (grayscale), and genome size. **e.** Number of novel vSGBs. The Venn diagram shows viruses shared between our skin vSGB set and external reference databases, as well as newly discovered viruses. **f.** Bars summarize the number of vSGBs with host matches at different taxonomic: cross-phylum, within phylum, within class, within order, within family, within genus, and single host.

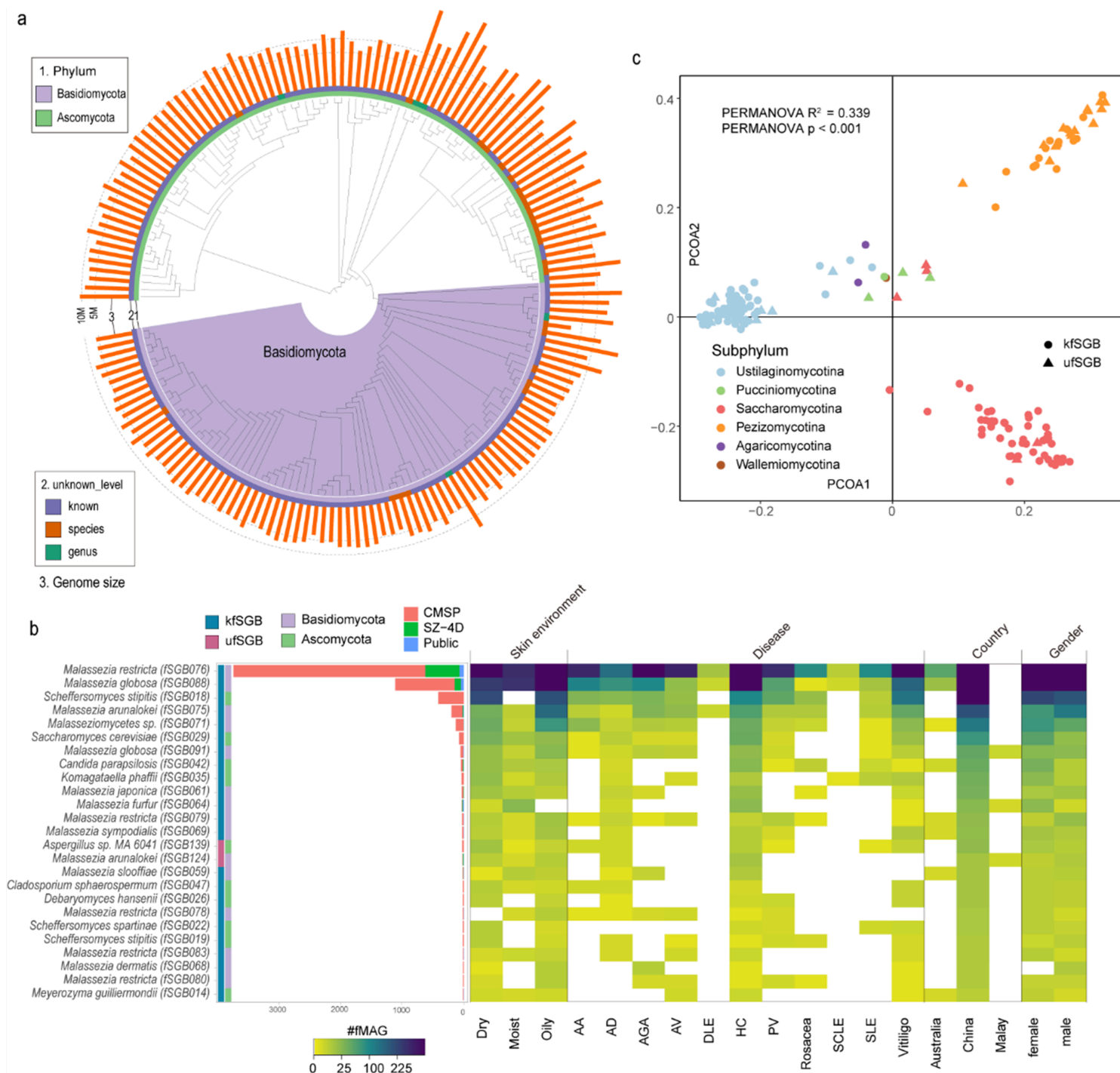


Figure 6

Diversity, composition, and ecology of skin fSGBs. **a.** Phylogenomic tree of representative fSGBs. Outer ring colors indicate phylum (Basidiomycota vs. Ascomycota). The second ring marks taxonomic novelty (known, genus-level, species-level unknown). Orange bars denote genome sizes. **b.** Distribution of fSGBs across skin environments and disease categories. In each panel, fSGBs are ordered by the number of fMAGs from the corresponding skin environment. For each fSGB, the total number of skin-environment fMAGs is reported, together with counts across 10 disease categories, 7 countries, and both genders. **c.** PCoA of fSGB functional profiles based on KEGG. Each point is an fSGB colored by phylum; triangles denote ufSGBs and circles denote kfSGBs.

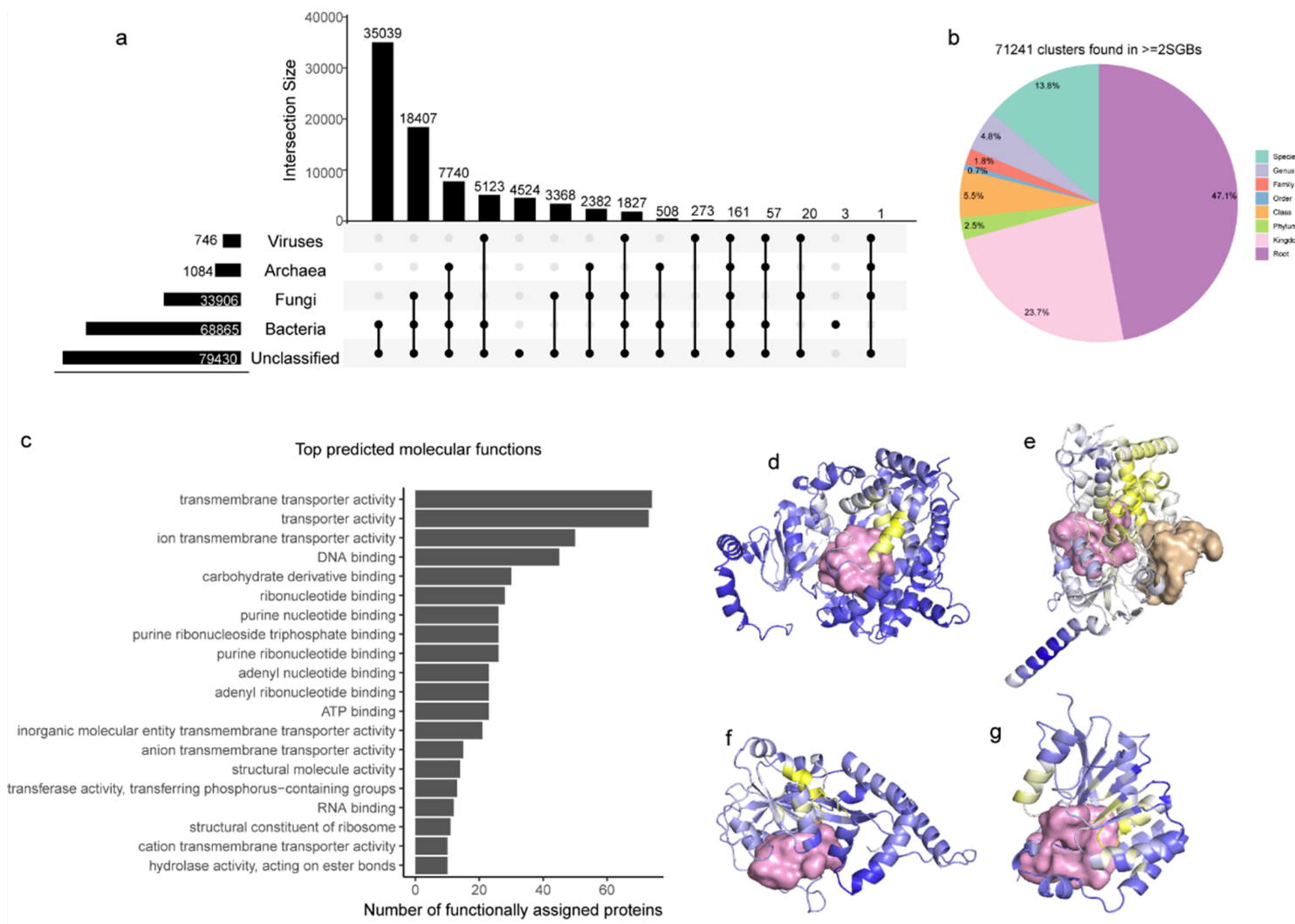


Figure 7

SNMPF cross-kingdom protein families and structure-based functional insights. **a.** The UpSet plot shows the domain-level taxonomic distribution of novel protein clusters. The total size of each category is indicated by the horizontal bar chart on the left. Intersections between categories are shown in the bottom matrix, and the sizes of these intersections are represented by the vertical bars at the top. **b.** Taxonomic breadth of SNMPF. Pie chart summarizes 71,241 clusters detected in at least two SGBs, grouped by the deepest shared taxonomic rank (species to root). **c.** Counts of the most frequently predicted GO molecular functions among the 808 functions predicted by DeepFRI. **d-g.** Examples of structures (d-e: families of monooxygenases, f-g: putative one-carbon transferases) with predicted pockets and functional annotations. Each example shows the UniProt ID (top), the highest-scoring DeepFRI function prediction (bottom) and the top-scoring pocket (pink surface). The structures are coloured by residue-level contributions to the DeepFRI function predictions, ranging from blue (no contribution) to yellow (strong contribution).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryMaterials20260511.docx](#)
- [Supplementarytables20260511.xlsx](#)