Skin Microbiome Signatures Identified via Non-Negative Matrix Factorization

Anastasia Ramanchanka
Modeling of Complex Biological System

June 25, 2025

Abstract

The skin microbiome is a complex and dynamic ecosystem whose disruption has been implicated in inflammatory skin conditions such as atopic dermatitis (AD) and psoriasis. While previous studies have often relied on broad taxonomic comparisons, these approaches fail to capture the ecological structure of microbial communities. In this study, I applied non-negative matrix factorization (NMF) to genus-level metagenomic data from 500 skin samples, including healthy individuals and patients with AD or psoriasis, to identify latent microbial configurations, or enterosignatures (ESs). Diversity analyses revealed significant beta diversity differences between AD and healthy samples but not for psoriasis. NMF decomposition identified seven enterosignatures, including a core signature dominated by *Cutibacterium acnes*, and disease-associated signatures enriched in *Micrococcus luteus* (AD) and *Staphylococcus epidermidis* (psoriasis). These findings highlight distinct ecological shifts associated with skin inflammation and demonstrate the utility of matrix factorization for uncovering compositional patterns beyond traditional clustering. Future work should explore causal links between microbial signatures and disease mechanisms using longitudinal and functional data.

1 Introduction

The skin acts as the body's main protective barrier, defending against environmental threats like pathogens and physical damage. It hosts a diverse microbial community that helps train the immune system, promoting tolerance to harmless microbes while fighting harmful ones. When this microbial balance is disrupted (dysbiosis), it can trigger inflammation, leading to skin conditions in genetically susceptible individuals. Many chronic diseases, including obesity [1], inflammatory bowel disease [2], psoriasis [3], allergic rhinitis [4], and atopic dermatitis [5], have been linked to microbial imbalances, though these connections are often correlations rather than proven causes.

The skin microbiome is a complex network of over 200 microbial species, including bacteria, fungi, viruses, and archaea [6]. It consists of both stable resident populations and temporary colonizers that may grow when the skin barrier is weakened [7]. Like the gut microbiome, skin microbes vary by body region [8], influenced by factors such as age [9], ethnicity [10], genetics [11], environment, and hygiene habits [12].

Simplifying the skin microbiome into key features is essential for medical research. Past studies used broad taxonomic groups (Bacteroidota vs. Firmicutes ratios) to link microbes to host traits [13]. However, these classifications ignore ecological relationships, such as bacteria that coexist or thrive under similar conditions.

In gut microbiome research, methods like ordination or clustering (e.g., PAM or DMM) group microbes into enterotypes (ETs), assuming each sample belongs to one distinct type [14]. Yet, these discrete groupings may overlook mixed bacterial communities. Alternative approaches, like latent Dirichlet allocation (LDA) or non-negative matrix factorization (NMF), analyze microbial ecosystems using continuous variables, capturing finer details [15] [16]. NMF is particularly useful because pre-defined bacterial signatures can be applied to individual samples, eliminating the need for large datasets required by ordination (e.g., PCA) or clustering (e.g., ET) methods.

Building upon previous work that decomposed the human gut microbiome using non-negative matrix factorization (NMF) into five microbial signatures ("enterosignatures" or ESs) [17], I applied a similar approach to analyze the skin microbiome.

2 Methods and Materials

2.1 Metagenomic dataset

This study uses publicly available skin microbiome data from the curatedMetagenomicData package, comprising 500 samples from individuals aged 20 to 80 years who had no recent antibiotic exposure. The dataset includes 374 samples from healthy individuals, 38 samples from individuals with atopic dermatitis (AD), and 88 samples from individuals with psoriasis.

2.2 Alpha and Beta diversity

To ensure robust comparisons and minimize confounding factors, alpha and beta diversity analyses were performed separately for each disease group (AD and psoriasis) versus healthy controls from the same study cohort. This approach was taken to avoid biases from regional differences in microbiome composition.

Alpha diversity measures the within-sample microbial diversity using Shannon index (H'), which considers both richness and evenness. It computes diversity metric from log-transformed counts and then t-test checks if the difference is statistically significant.

Beta Diversity measures between-sample microbial variation, showing how different microbial communities are from each other. Bray-Curtis dissimilarity was used, which compares the abundances of species that are shared between two samples, and the number of species found in each. To visualize, principal coordinate analysis (PCoA) was applied, which transforms the multidimensional distance data into an interpretable two-dimensional representation.

In order to establish which species are differentially abundant between disease groups and healthy controls, ANCOM-BC (Analysis of Composition of Microbiomes with Bias Correction) was employed. The method first converted the data into a phyloseq object and analyzed relative abundance data while accounting for compositionality bias. Statistical testing was performed using a linear model with disease status as the fixed effect. Results were filtered for significant taxa (absolute log-fold change > 0 and -log10 q-value > 5). The findings were visualized through a volcano plot displaying effect sizes (β coefficients) against statistical significance (-log10 q-values), with differentially abundant taxa highlighted and labeled.

2.3 Non-negative matrix factorisation

Genus-level relative abundance matrices were normalized by the sum of abundances per sample after low-prevalence taxa were excluded (observed in fewer than 10% of samples) and samples with insufficient taxonomic representation (containing fewer than 10% of prevalent species). NMF was performed with Scikit-Learn v0.24.167 using the multiplicative update solver, Kullback-Leibler divergence as a beta-loss function, random initialisation, and a maximal number of iterations of 2000. All runs were performed for each number of clusters k, ranging from 2 to 20. The quality of decomposition for the validation was calculated with the explained variance:

$$EV = 1 - \frac{\sum (x_{ij} - \widehat{x_{ij}})^2}{\sum x_{ij}^2} \tag{1}$$

The higher the EV value, the more accurately the decomposition represents the input compositional matrix X.

The quality of the decomposition was further estimated using a cosine similarity between the reconstructed matrix and the original abundance matrix, or a sample microbial composition profile and its prediction:

$$CS = \frac{\sum x_i \widehat{x_i}}{\sqrt{\sum x_i^2}\sqrt{\sum \widehat{x_i^2}}} \tag{2}$$

The optimal number of signatures was chosen by assessing the gain of cosine similarity when increasing the number of signatures: when the significant gain no longer increases, the optimal number of signatures is reached. To objectively identify this inflection point, the KneeLocator package was employed, which detects the maximum curvature (knee point) in the similarity improvement curve [18]. Given the noise and non-linearity in the curve, I applied polynomial interpolation to better reveal the underlying trend. While first-degree (linear) and second-degree (quadratic) polynomials failed to capture the knee point, higher-degree polynomials (4th and 5th) provided better fits. To balance accuracy and avoid overfitting, I selected a fifth-degree polynomial with knee point at k=7.

2.4 Enterosignature analyses

The matrices H and W resulting from the NMF algorithm represent the weight of genera in enterosignatures and the presence of enterosignatures in samples. Normalizing H by its columns or its rows informs on the general composition of ESs in genera, and on the association strength of genera to each ES, respectively. By normalizing the W matrix column-wise, the relative abundance of ES in each sample is obtained.

3 Results

3.1 Alpha and Beta diversity

Alpha diversity metrics reveal no statistically significant differences in within-sample microbial diversity among healthy, AD, and psoriasis groups, suggesting that bacterial community richness and evenness remain stable regardless of disease status (Figure 1).

Beta diversity analysis using Bray-Curtis dissimilarity revealed distinct patterns between groups. While no significant clustering was observed between healthy and psoriasis samples

Figure 1: Alpha Diversity Across Health Conditions: Atopic Dermatitis (AD) vs. Healthy Samples (left) and Psoriasis vs. Healthy Samples (right)

on the PCoA plot (Figure 2), suggesting similar overall microbiome composition, a clear separation emerged between atopic dermatitis (AD) and healthy groups. This divergence indicates that AD is associated with specific microbial community features that differentiate it from healthy skin.

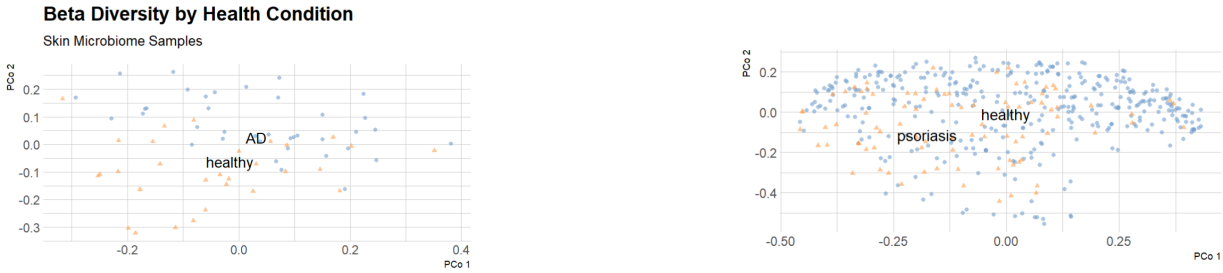

Figure 2: Beta Diversity Across Health Conditions: Atopic Dermatitis (AD) vs. Healthy Samples (left) and Psoriasis vs. Healthy Samples (right)

Where beta diversity analysis revealed distinct clustering between AD and healthy skin samples, I next identified differentially abundant taxa using ANCOM-BC, a compositional methodology accounting for microbiome data constraints. For AD samples (threshold: absolute $\log_2$-fold change > 1, $q < 10^{-5}$), *Corynebacterium amycolatum* and *Gordonia bronchialis* were significantly enriched, while *Anaerococcus octavius* was more abundant in healthy controls. In psoriasis samples, only *Cutibacterium acnes* met the same stringent thresholds, though relaxing the fold-change criterion to > 0.7 additionally implicated *Malassezia restricta* (Figure 3). Notably, these observational findings demonstrate statistical associations but do not establish causal relationships or ecological interactions between these species and disease states. Further non-negative matrix factorization (NMF) analysis is required to validate their biological relevance.

3.2 Finding optimal number of signatures

The knee locator algorithm identified the point of maximum curvature at $k = 7$ signatures. As shown in Figure 4 (top panel), the empirical curve exhibits substantial noise and nonlinearity, where the true knee point is obscured by stochastic fluctuations. To improve robustness, I applied polynomial interpolation to smooth the curve.

Figure 3: Volcano Plot of Differentially Abundance Bacterial Species Across Health Conditions: Atopic Dermatitis (AD) vs. Healthy Samples (left) and Healthy Samples vs Psoriasis (right)

While first-degree (linear) and second-degree (quadratic) polynomials failed to capture knee points in the trend, higher-order polynomials (degrees 4-6) provided better fits. However, excessive polynomial degrees risk overfitting. I selected a fifth-degree polynomial, which results in the same number of $k = 7$ signatures.

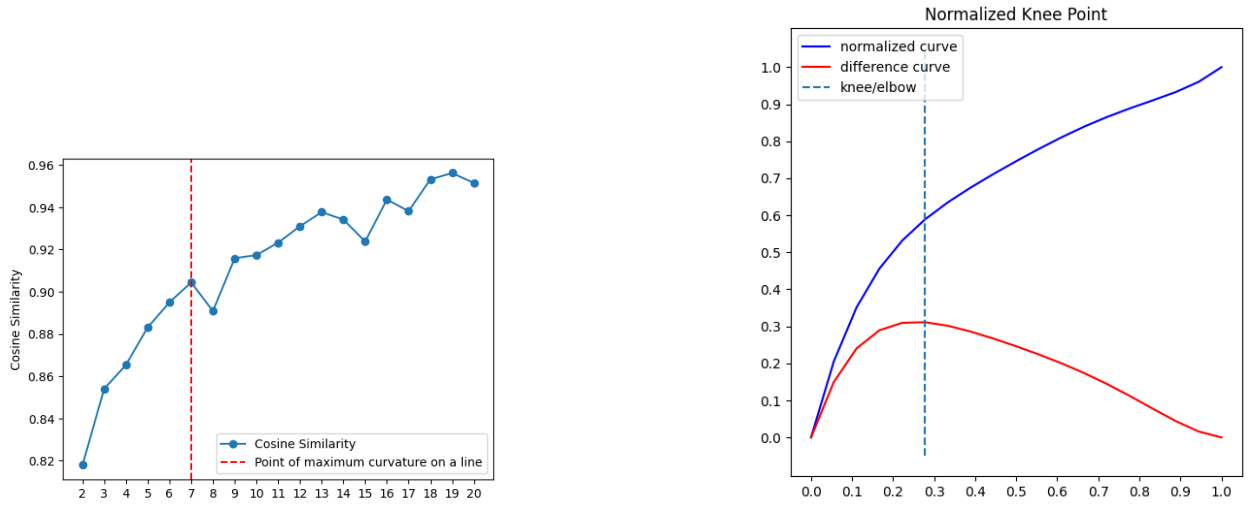

Figure 4: Cosine similarity versus number of enterosignatures (k). The point of maximum curvature ($k = 7$, dotted line) identifies the optimal signature count.

3.3 Enterosignature analyses

Subsequent analyses were performed using the optimal number of enterosignatures (k=7). Initial examination focused on the distribution of signature contributions across samples, visualized in the clustered heatmap (Figure 5).

Notably, while most enterosignatures (1-6) showed strong, sample-specific associations, each being predominantly represented in distinct sample clusters, ES 7 demonstrated a markedly different pattern. This signature was nearly uniformly distributed across the majority of samples, suggesting it may represent a core microbial community component common to most samples.

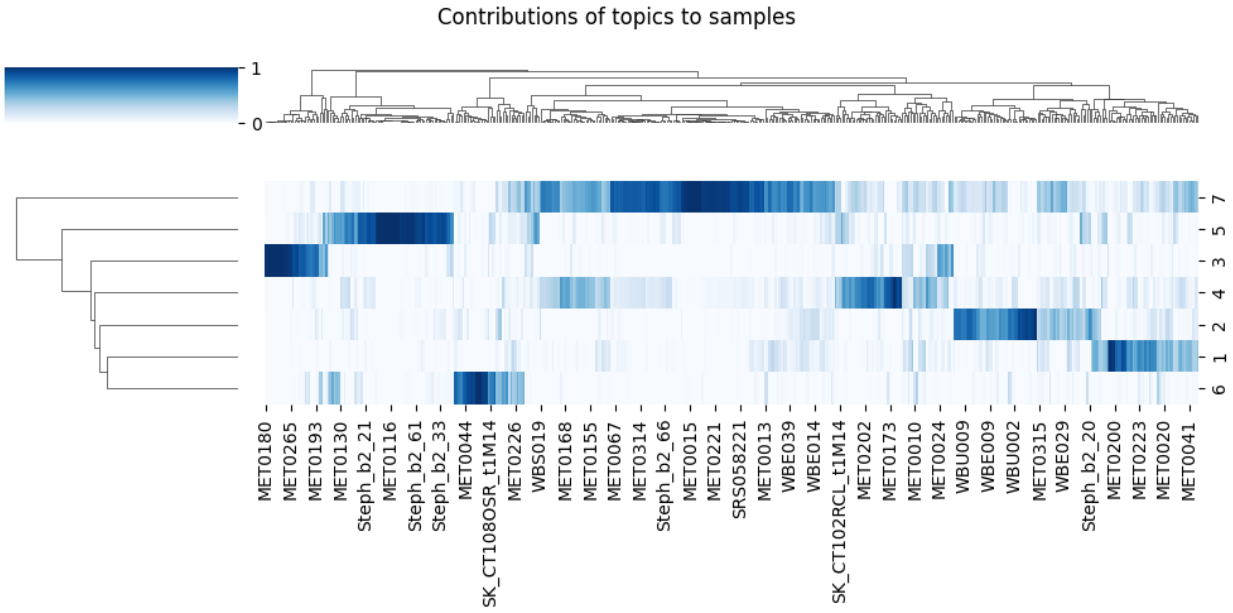

Figure 5: Enter Caption

To characterize the taxonomic composition of each enterosignature, I generated a heatmap visualizing the top five most abundant species per topic. Distinct patterns emerged among the seven topics: Topics 1-6 each demonstrated unique species assemblages, featuring dominant taxa such as *Staphylococcus hominis* in Topic 3, *Malassezia restricta* in Topic 4, and *Staphylococcus epidermidis* in Topic 5. In striking contrast, Topic 7 showed exceptional specificity, with *Cutibacterium acnes* comprising 98.3% of its composition. This marked disparity suggests Topic 7 represents either a nearly monocultural microbial state and universal skin colonizer fundamentally distinct from the niche-specific signatures observed in Topics 1-6.

Box plot visualization of enterosignature contributions across disease states (Figure 7) revealed Topic 7 maintained its pan-sample distribution, consistent with a ubiquitous baseline signature, while disease-specific signatures emerged: Topic 2 showed significant enrichment in AD samples versus controls and contained AD-associated *Micrococcus luteus*, whereas Topic 5 dominated psoriasis samples with *Staphylococcus epidermidis*, suggesting distinct microbial ecologies underlie these inflammatory conditions.

While signature contributions were examined across age groups (Figure 8), limited sample size precluded robust statistical interpretation. Future studies with balanced cohorts are needed to assess age-related effects.

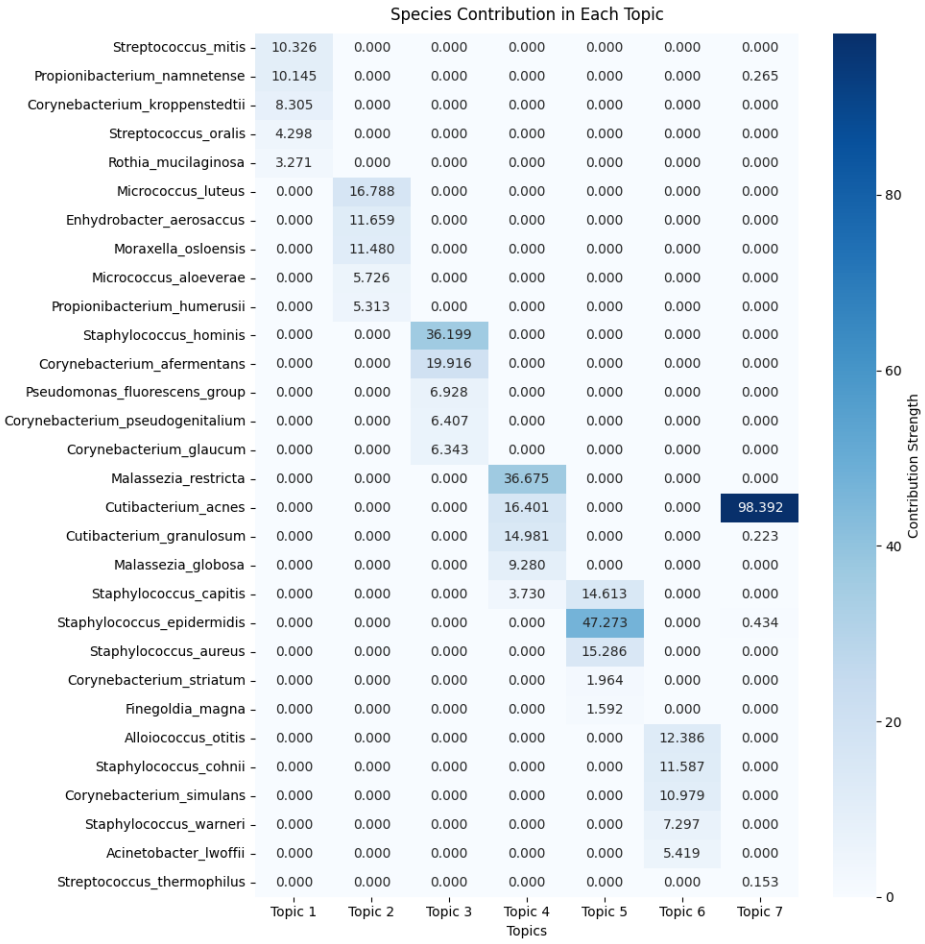

Figure 6: Species Contribution Across Topics (%)

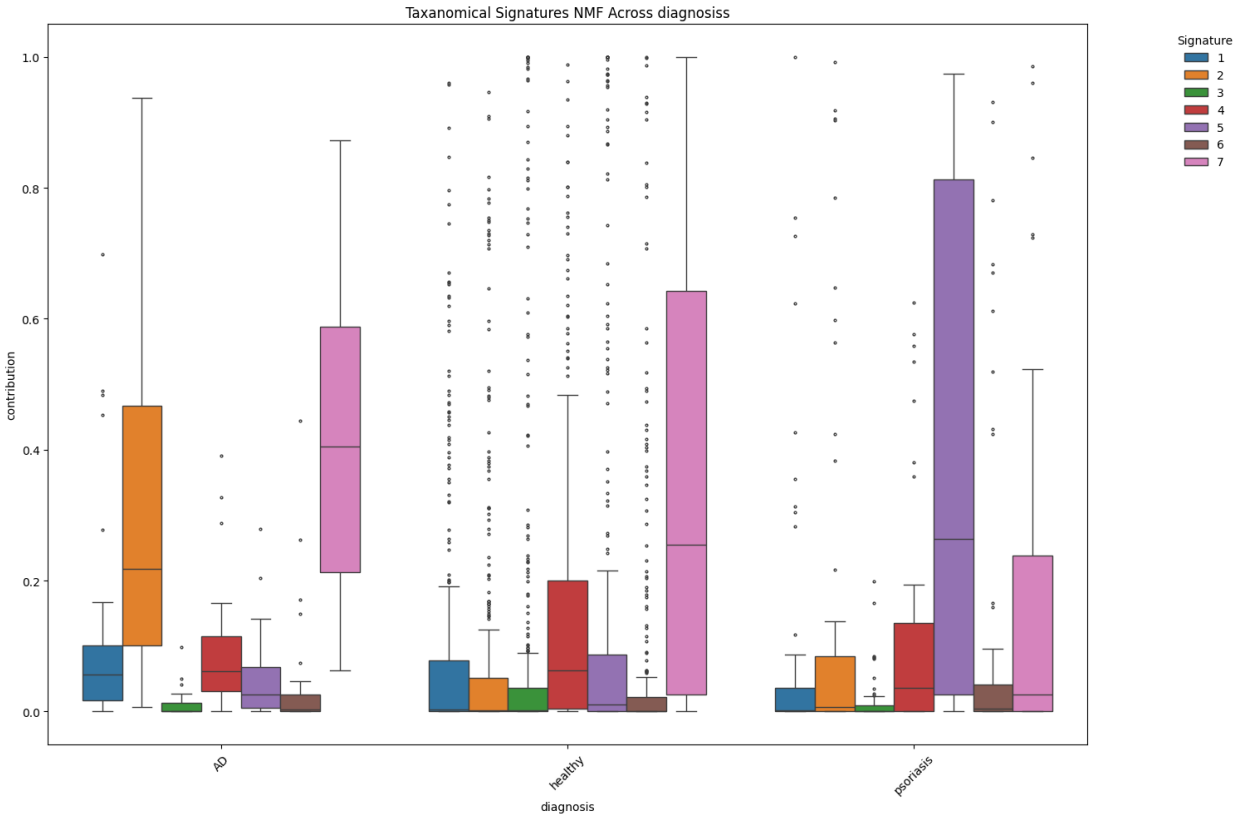

Figure 7: Enterosignature Contributions by Health Status

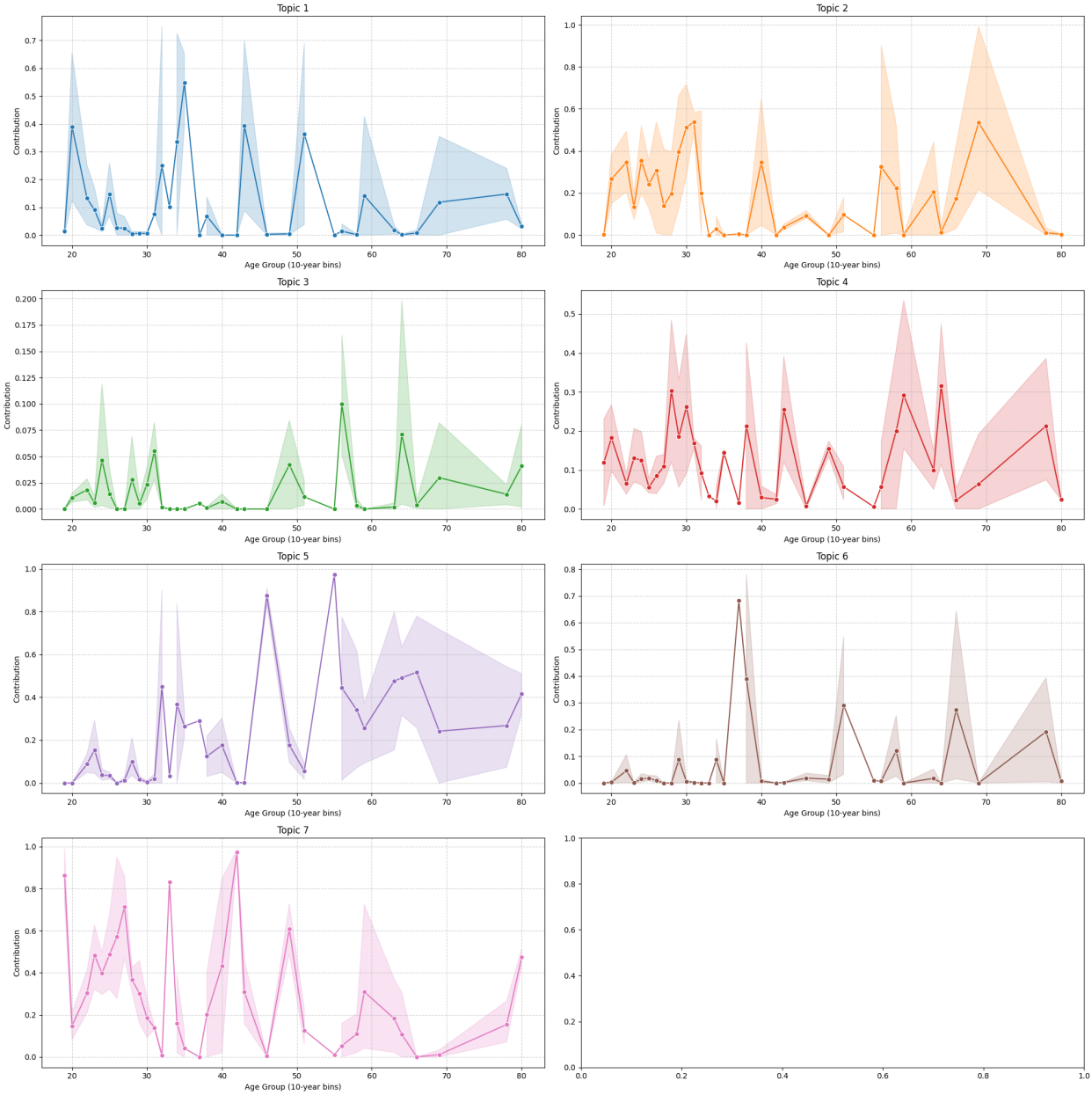

Figure 8: Age Distribution of Enterosignature Contributions

4 Discussion

This study applies non-negative matrix factorization (NMF) to decompose the skin microbiome into seven enterosignatures (ESs), revealing both conserved and disease-associated microbial configurations. The findings align with established ecological principles of skin microbial communities while identifying novel compositional patterns linked to dermatopathological states.

ES7 exhibited near-universal distribution across samples, dominated by *Cutibacterium acnes* (98.3%). This aligns with its recognized role as a commensal skin colonizer, where its depletion has been associated with dysbiosis in acne, atopic dermatitis (AD), and psoriasis. The ubiquity of ES7 suggests it represents a foundational component of the skin microbiome, resilient to perturbations that alter niche-specific communities.

ES2 was significantly enriched in AD samples and characterized by *Micrococcus luteus*, a taxon previously implicated in AD pathogenesis due to its pro-inflammatory potential. Conversely, ES5 dominated psoriasis samples and was primarily composed of *Staphylococcus epidermidis*. While *S. epidermidis* is typically commensal, its overrepresentation in psoriasis may reflect competitive exclusion of *S. aureus*, a modulator of cutaneous inflammation [8]. These signatures suggest distinct ecological disruptions underlying each disease, though mechanistic validation is required to distinguish causation from correlation.

The NMF-derived signatures captured both discrete and gradient-like microbial distributions, overcoming limitations of traditional clustering (enterotypes). However, the study's observational design precludes causal inference. Longitudinal cohorts and metatranscriptomic data could clarify whether signature shifts drive pathology or result from inflammatory microenvironments.

Future directions would be link ESs to host immune markers via multi-omics integration; Metagenomic assembly could clarify if S. epidermidis in ES5 represents commensal or pathogenic strains.

References

1. Torres-Fuentes, C., Schellekens, H., Dinan, T. & Cryan, J. The microbiota-gut-brain axis in obesity. *Lancet Gastroenterol Hepatol* (2017).
2. Sheehan, D. & Shanahan, F. The Gut Microbiota in Inflammatory Bowel Disease. *Gastroenterol Clin North Am* (2017).
3. Celoria, V. *et al.* The Skin Microbiome and Its Role in Psoriasis: A Review. Psoriasis. *Auckland, N.Z.* (2023).
4. Hyun, D. *et al.* Dysbiosis of Inferior Turbinate Microbiota Is Associated with High Total IgE Levels in Patients with Allergic Rhinitis. *Infect. Immun* (2018).
5. Yamazaki, Y., Nakamura, Y. & Nunez, G. Role of the microbiota in skin immunity and atopic dermatitis. *Allergol. Int.* (2017).
6. Cundell, A. Microbial Ecology of the Human Skin. *Microb. Ecol.* (2018).
7. Kong, H. Temporal shifts in the skin microbiome associated with disease flares and treatment in children with atopic dermatitis. *Genome Res* (2012).

8. Oh, J. *et al.* Comparative Sequencing Program Biogeography and individuality shape function in the human skin metagenome. *Nature* (2014).

9. Costello, E. K. *et al.* Bacterial community variation in human body habitats across space and time. *Science* **326,** 1694–1697 (2009).

10. Gupta, V. K., Paul, S. & Dutta, C. Geography, Ethnicity or Subsistence-Specific Variations in Human Microbiome Composition and Diversity. *Frontiers in Microbiology* **8,** 1162 (2017).

11. Kim, H. J. *et al.* Segregation of age-related skin microbiome characteristics by functionality. *Scientific Reports* **9,** 16748 (2019).

12. Staudinger, T., Pipal, A. & Redl, B. Molecular analysis of the prevalent microbiota of human male and female forehead skin compared to forearm skin and the influence of make-up. *Journal of Applied Microbiology* **110,** 1381–1389 (2011).

13. Ley, R., Turnbaugh, P., Klein, S. & Gordon, J. Microbial ecology: human gut microbes associated with obesity. *Nature* (2006).

14. Arumugam, M. & et al. Enterotypes of the human gut microbiome. *Nature* (2011).

15. Cai, Y., Gu, H. & Kenney, T. Learning Microbial Community Structures with Supervised and Unsupervised Non-negative Matrix Factorization. *Microbiome* (2017).

16. Breuninger, T. & et al. Associations between habitual diet, metabolic disease, and the gut microbiota using latent Dirichlet allocation. *Microbiome* (2021).

17. Frioux, C. & et al. Enterosignatures define common bacterial guilds in the human gut microbiome. *Microbiome* (2023).

18. Satopaa, V., Albrecht, J., Irwin, D. & Raghavan, B. "Finding a "Kneedle" in a Haystack: Detecting Knee Points in System Behavior. *31st International Conference on Distributed Computing Systems Workshops* (2011).