Breast Milk Microbiome Shifts from Environmental to Infant-Oral Signatures Across Lactation Stages

Dalia H. Almasud

Mentors: Lajos Kalmar, Hugo Tavares

Abstract

Human breast milk contains a living bacterial community whose composition is not fixed. This project investigated whether that composition changes systematically across the postpartum period, and whether such change reflects a biologically coherent shift consistent with retrograde inoculation, the transfer of bacteria from an infant's developing oral microbiome back into milk during nursing. Using 16S rRNA sequencing data from 540 breast milk samples spanning the first several years postpartum (Qiita Study 11366, Chambers Breast Milk Microbiome), samples were grouped into four clinical lactation stages and analyzed for diversity, differential abundance, and compositional structure. Microbial diversity increased significantly across stages (Kruskal-Wallis, $p=2.7\times10^{-7}$), and community composition also differed significantly by stage (PERMANOVA, $R^2=1.82\%$, $p=0.001$). Twenty-three genera changed significantly with stage (DESeq2, padj<0.1): oral-associated genera consistently increased while environmental and enteric genera consistently declined, a pattern independently cross-validated with a second, compositional-data-based method (ALDEx2, 91% overlap). One ambiguous taxon, a SILVA placeholder cluster, was resolved via BLAST to a dominant sequence matching Prevotella melaninogenica at 100% identity, a documented infant oral colonizer. Together, these results show that breast milk composition shifts consistently over the postpartum period toward genera linked to the infant oral microbiome, a pattern consistent with, though not proof of, retrograde inoculation.

Dataset: Qiita Study 11366 — https://qiita.ucsd.edu/study/description/11366

1. Introduction

Human breast milk contains a living bacterial community whose composition is not static. It may shift over time as an infant's own oral microbiome develops, potentially through the backflow of infant saliva into the breast during nursing, a process known as retrograde inoculation (Moossavi & Azad, 2020). Prior work has established that milk composition changes over the course of lactation (Hunt et al., 2011; Cabrera-Rubio et al., 2012), and that lactation stage specifically, rather than birth mode, is a primary driver of milk microbiome variation (Lyons et al., 2022). However, the specific genus-level succession pattern across finely resolved postpartum stages, and whether it aligns with genera independently documented as infant oral colonizers, has not been comprehensively characterized in a single, cross-validated analysis.

Problem statement: This project asks whether breast milk microbial diversity and community composition change systematically across postpartum time, and whether that change is concentrated in genera linked to the developing infant oral microbiome.

Key assumptions and limitations: 16S rRNA amplicon sequencing identifies bacteria at approximately the genus level from a single short gene region; it cannot resolve full species identity in all cases, nor can it establish the direction of bacterial transfer between mother and infant. These limitations are addressed directly in Section 6.

Summary of approach: 540 breast milk samples were grouped into four clinical lactation stages and analyzed using alpha diversity, beta diversity (PERMANOVA), and genus-level differential abundance testing (DESeq2), cross-validated with an independent compositional-data method (ALDEx2) and several robustness checks (contamination screening, delivery-mode adjustment, outlier sensitivity analysis). Two ambiguous genus-level identities were resolved via BLAST against NCBI reference sequences.

2. Related Work

Several studies have characterized how the human milk microbiome changes over lactation. Hunt et al. (2011) first characterized the diversity and temporal stability of milk bacterial communities. Cabrera-Rubio et al. (2012) showed milk composition changes over lactation and is shaped by maternal weight and delivery mode. Lyons et al. (2022) found that milk microbiome composition aligns with lactation stage rather than birth mode, consistent with the present project's own robustness finding that delivery mode does not explain the observed stage effect. Notably, Lyons et al. reported milk diversity decreasing across lactation, the opposite direction from the present findings; this discrepancy is addressed in Section 5. Moossavi & Azad (2020) reviewed proposed origins of human milk microbiota, including retrograde inoculation, the biological hypothesis this project's genus-level pattern is consistent with. Filatava et al. (2023) examined how gestational age and postpartum week jointly shape milk microbiota in preterm infants, directly relevant to this project's own investigation of two preterm-associated outlier samples (Section 5).

Methodologically, this project builds on standard tools with established citations: DADA2 for amplicon sequence variant inference (Callahan et al., 2016), the SILVA reference database for taxonomy (Quast et al., 2013), phyloseq for data structuring (McMurdie & Holmes, 2013), DESeq2 for differential abundance testing (Love et al., 2014), ALDEx2 for compositional-data cross-validation (Fernandes et al., 2014), and decontam for contamination screening (Davis et al., 2018). Douglas et al. (2020) and Salter et al. (2014) motivate this project's contamination and extraction-related robustness checks.

Gap addressed: Prior milk microbiome studies typically report a single differential abundance method without independent statistical cross-validation, and rarely resolve ambiguous or database-placeholder taxonomic assignments to the sequence level. This project combines four convergent lines of evidence (alpha diversity, beta diversity, DESeq2, and ALDEx2 cross-validation) with direct BLAST-based species resolution for ambiguous genera, an integrated approach not present in the cited prior work.

3. Methodology

3.1 Method Details

The analysis pipeline proceeded through nine stages (Figure 1):

- Data source: 540 breast milk samples, 16S rRNA gene, V4 region (primers 515F/806R), sequenced on Illumina MiSeq (Qiita Study 11366).
- Sequence processing: Raw reads denoised and resolved into amplicon sequence variants (ASVs) using DADA2 (Callahan et al., 2016).
- Taxonomy assignment: ASVs classified against the SILVA v138.1 reference database (Quast et al., 2013).
- Data structure and filtering: Data merged into a phyloseq object (McMurdie & Holmes, 2013); ASVs collapsed to genus level via tax_glom; antibiotic-exposed samples excluded. Final dataset: 540 samples, 799 genera.
- Clinical staging: Days postpartum binned into four stages following the standard clinical transitional-to-mature milk framework (Pang & Hartmann, 2007; Ballard & Morrow, 2013): Transitional (5–14 days), as milk shifts from colostrum toward mature composition; and three finer subdivisions applied within the mature-milk period for temporal resolution across this dataset's wide postpartum range: Early mature (15–90 days), Established mature (91–365 days, coinciding with tooth eruption and solid-food introduction), and Extended mature (>365 days).
- Diversity analysis: Alpha diversity (Shannon index, via phyloseq) and beta diversity (Bray-Curtis PERMANOVA, via vegan's adonis2()).
- Differential abundance: DESeq2 (Love et al., 2014), modeling genus counts as a function of sequencing batch and lactation stage. A likelihood ratio test (LRT) provided the omnibus significance test across all four stages; explicit pairwise contrasts described direction and magnitude for specific stage transitions.
- Cross-validation: ALDEx2 (Fernandes et al., 2014), an independent compositional (CLR-transformed) method with Monte Carlo resampling, run on the same data as a statistical sanity check against DESeq2.

- Robustness checks: decontam (Davis et al., 2018) using 122 negative control samples; delivery mode added as a covariate; outlier sensitivity analysis (PERMANOVA re-run with and without two identified high-diversity outlier samples).

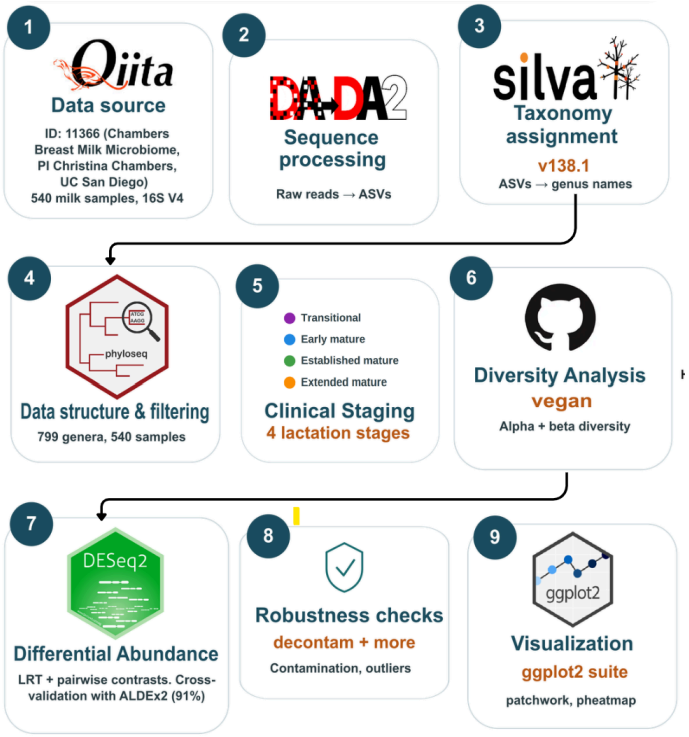


Figure 1. The nine-stage analysis pipeline, from data source through visualization.

3.2 Evaluation Procedure

Tools, libraries, and frameworks: R (DADA2, phyloseq, DESeq2, ALDEx2, vegan, decontam, ggplot2, patchwork, pheatmap); NCBI BLASTN for sequence-level species confirmation.

Dataset: Qiita Study 11366 (Chambers Breast Milk Microbiome, UC San Diego), a public 16S rRNA amplicon dataset; no synthetic data was used.

Metrics and significance criteria: Genus-level significance was assessed with DESeq2's LRT (padj<0.1, Benjamini-Hochberg corrected). Diversity significance used Kruskal-Wallis (alpha diversity, categorical stages) and Spearman correlation (continuous postpartum time); Dunn's test with BH correction for post-hoc pairwise comparisons. Community composition was assessed via PERMANOVA R^2 and p-value. The cross-validation criterion was the proportion of ALDEx2-significant genera also present in the DESeq2 LRT list. Taxonomic resolution used BLAST percent identity and query coverage against NCBI 16S reference sequences.

4. Results

Baseline consideration: As no prior study analyzing this specific dataset exists for direct comparison, results are contextualized against the broader milk microbiome literature (Section 2) and against internal robustness checks (with/without covariate adjustment, with/without outlier samples) as an effective baseline for judging result stability.

Alpha Diversity

Shannon diversity increased significantly across the four lactation stages (Kruskal-Wallis, $p=2.7\times10^{-7}$; Figure 2). Post-hoc Dunn's testing (BH-adjusted) showed this increase was concentrated in later-stage comparisons, with the strongest separation between Early and Extended mature milk ($p<0.0001$); the earliest transitions (Transitional vs. Early or Established mature) were not statistically significant, plausibly reflecting the small Transitional group size (n=13).

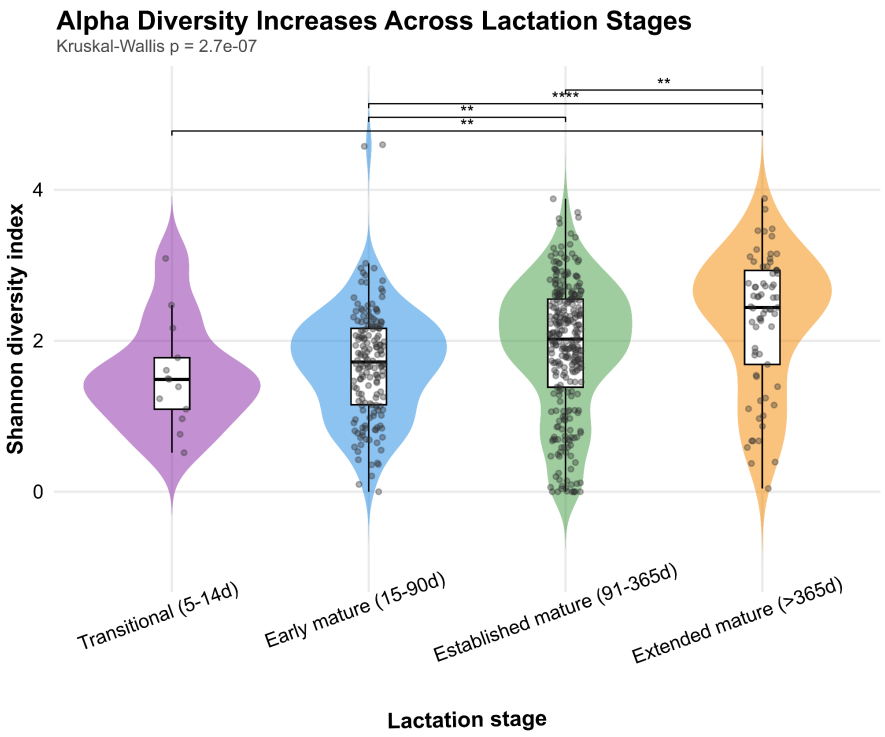


Figure 2. Shannon diversity index by clinical lactation stage.

Beta Diversity

Community composition also differed significantly by stage (PERMANOVA, R^2=1.82%, p=0.001). Effect size was modest in absolute terms, consistent with the well-established observation that individual-level variation dominates human microbiome community composition (typically 90–98% of total variance).

Differential Abundance

The DESeq2 LRT identified 23 genera significantly associated with lactation stage (padj<0.1; Figure 3). Across four pairwise stage comparisons, oral-associated genera (e.g., Granulicatella, Neisseria, Haemophilus, Veillonella, Rothia) consistently increased, while environmental and enteric genera (e.g., Acinetobacter, Pseudomonas, Klebsiella) consistently declined, most sharply by the Extended mature stage.

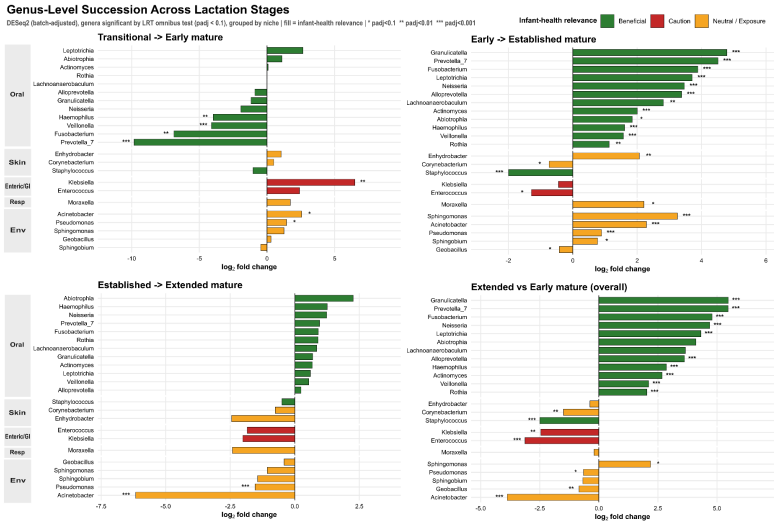


Figure 3. Genus-level log2 fold change across four pairwise lactation-stage comparisons, colored by infant-health relevance tier.

Individual-Sample Confirmation

A heatmap of the same 23 genera across all 540 individual samples (Figure 4), row z-scored against each genus's own dataset-wide mean, confirmed the succession pattern holds at the individual-sample level, not only as a group-average artifact.

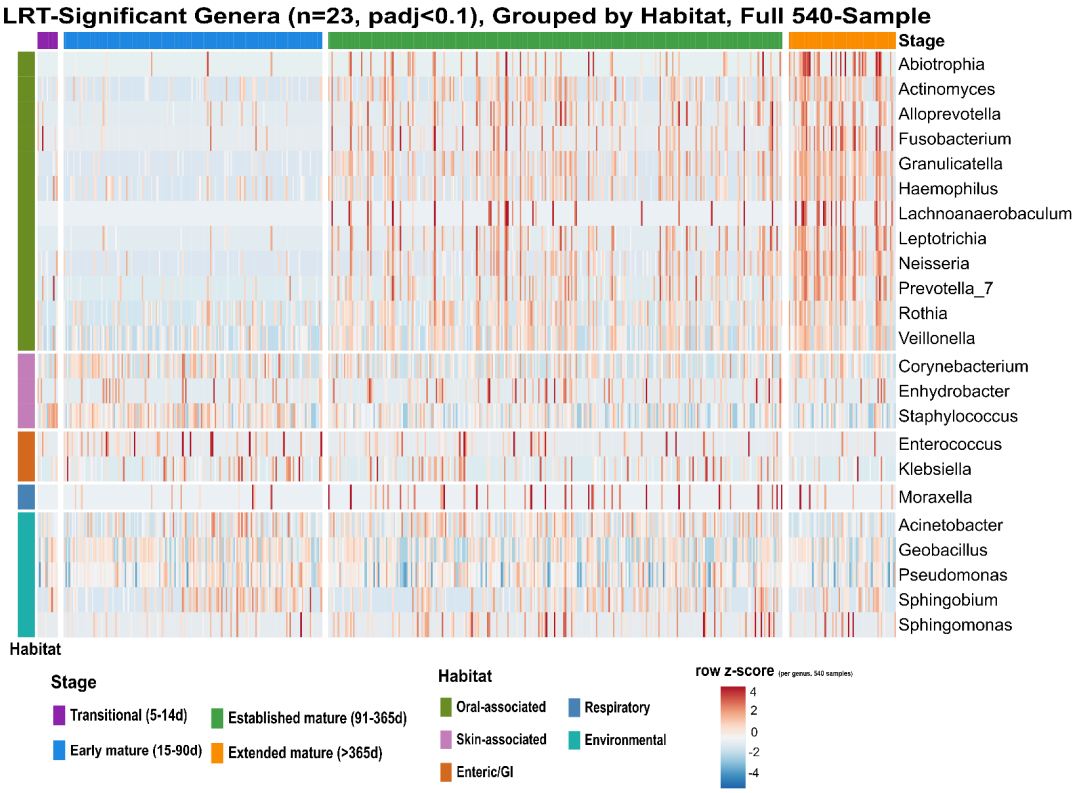


Figure 4. Row z-scored abundance of the 23 significant genera across all 540 individual samples, grouped by habitat category.

Infant-Colonizer-Specific Trend

Eight genera meeting both statistical significance and independent literature evidence for infant oral colonization showed a shared dip at Early mature followed by a steady rise through Established and Extended mature (Figure 5).

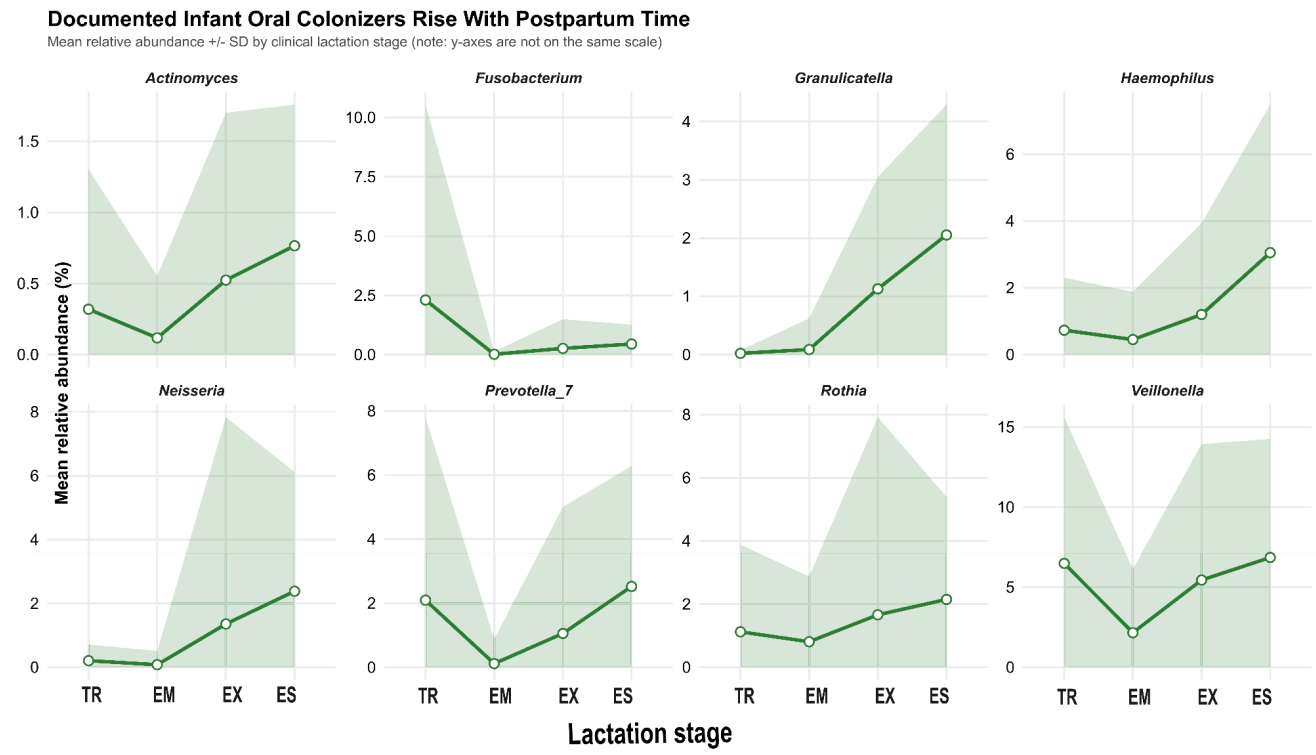


Figure 5. Mean relative abundance trajectories for eight documented infant oral colonizer genera across lactation stages.

Cross-Validation and Robustness

ALDEx2 independently identified 11 significant genera; 10 of these (91%) overlapped with the DESeq2 LRT list. The core PERMANOVA result was materially unchanged after removing two high-diversity outlier samples ($p=0.001$, $R^2 \approx 1.8\%$ with or without). Preterm birth status, tested as a covariate, was not a significant independent driver of community composition once lactation stage and technical batch were controlled for.

Taxonomic Resolution

Prevotella_7, a SILVA placeholder for an unresolved sequence cluster, was resolved via BLAST: its dominant sequence variant (>90% of the cluster's reads) matched Prevotella melaninogenica at 100% identity and full query coverage, a documented infant oral colonizer.

Method	Result	Significance
Alpha diversity (Kruskal-Wallis)	Shannon diversity rises with stage	$p = 2.7 \times 10^{-7}$
Beta diversity (PERMANOVA)	$R^2 = 1.82\%$	$p = 0.001$
Differential abundance (DESeq2 LRT)	23 significant genera	padj < 0.1
Cross-validation (ALDEx2)	91% overlap with DESeq2	10/11 genera

Table 1. Summary of the four convergent lines of evidence for the lactation-stage effect.

5. Discussion

The convergence of four independent analytical approaches, each with different statistical assumptions, on the same directional pattern is the strongest evidence in this project that the lactation-stage effect on milk microbiome composition is real rather than a modeling artifact. This is particularly notable given that DESeq2

(negative-binomial, size-factor normalized) and ALDEx2 (compositional, CLR-transformed with Monte Carlo resampling) make substantially different assumptions about the underlying data-generating process, yet converge on 91% of significant genera.

The biological interpretation, that oral-associated genera increase as environmental and skin-associated genera decline, is consistent with retrograde inoculation: as an infant's own oral microbiome matures (coinciding with tooth eruption and solid-food introduction around the Established mature stage), a growing and increasingly complex source of oral bacteria may be introduced back into the breast during nursing. This also offers a plausible account for the dip-then-rise trajectory observed among documented infant colonizers (Figure 5): residual birth-related or environmental seeding may dominate the Transitional stage, fading by Early mature before the infant's own maturing oral microbiome begins driving the later rise.

One unexpected finding concerns Streptococcus: despite being one of the most well-documented infant oral colonizers in the literature and the fourth most abundant genus overall (12.45% of all reads), it narrowly missed the DESeq2 significance threshold (padj=0.207), while a simpler Kruskal-Wallis test on its raw relative abundance was highly significant (p=0.0002). This divergence reflects Streptococcus's high count variance as a dominant taxon, a statistical property of the stricter DESeq2 test rather than an absence of real biological signal.

A second unexpected finding involved two high-diversity outlier samples in the Early mature stage, both of which were preterm births, a striking coincidence given preterm status represents a minority of the dataset (11–13%). Formal investigation found this outlier diversity was driven by single-genus hyper-dominance (a "bloom") combined with sequencing batch, not a true biological subgroup; a subsequent multivariable PERMANOVA confirmed preterm status was not an independent driver of composition once stage and batch were controlled for. This result also partially diverges from Lyons et al. (2022), who reported milk diversity decreasing across lactation, the opposite direction from the present findings; the two studies agree, however, that lactation stage, not birth mode, is the primary compositional driver, suggesting the diversity-direction discrepancy may reflect differences in cohort, geography, or the specific stage boundaries used, rather than a methodological error in either study.

6. Limitations and Future Work

- Limitation: 16S rRNA data cannot establish the direction of bacterial transfer between mother and infant. Next step: shotgun metagenomics or infant oral microbiome co-sampling would allow direct strain-level tracking to test transfer directionality.
- Limitation: No true colostrum-stage samples were available in this dataset (earliest sample: day 5.1 postpartum). Next step: a cohort including colostrum-stage sampling would extend the succession timeline to its true starting point.
- Limitation: Prevotella_7 remains an unresolved cluster of closely related species rather than a single confirmed organism, even after BLAST resolution of its dominant sequence variant. Next step: full-length 16S or shotgun sequencing would enable definitive species-level resolution of the entire cluster, not just its dominant member.
- Limitation: Beta diversity effect size is modest in absolute terms (R^2=1.82%), as expected for human microbiome data where individual variation dominates. Next step: a larger, longitudinal within-subject cohort (repeated sampling of the same mother-infant pairs over time) would increase power to detect within-individual stage effects specifically.
- Promising future direction: formally decompose the observed Shannon diversity increase into its richness and evenness components, since both increased detection of new genera and more even distribution among existing genera could plausibly contribute to the observed pattern, and this analysis does not distinguish between them.
- Promising future direction: a visually distinct, unexamined sub-cluster of Established mature samples was noted during ordination analysis but not investigated further given time constraints; characterizing this sub-cluster could reveal an additional, currently unexplained biological subgroup.

7. Conclusion

This project shows that breast milk microbial composition shifts consistently across the postpartum period, moving away from environmental and skin-associated bacteria and toward genera with documented roles in the infant oral microbiome. This pattern was supported by four convergent, independent lines of evidence, alpha diversity, beta

diversity, DESeq2 differential abundance testing, and ALDEx2 cross-validation, and survived multiple robustness checks, including contamination screening, delivery-mode adjustment, and outlier sensitivity analysis. One ambiguous taxon was resolved to a documented infant oral colonizer via direct BLAST confirmation. Together, these results are consistent with retrograde inoculation of milk via infant oral contact during nursing, though 16S data alone cannot establish transfer direction definitively, the main limitation of this analysis and the clearest direction for future work.

References

1. Ballard, O., & Morrow, A. L. (2013). Human milk composition: nutrients and bioactive factors. Pediatric Clinics of North America, 60(1), 49–74.
2. Cabrera-Rubio, R., Collado, M. C., Laitinen, K., Salminen, S., Isolauri, E., & Mira, A. (2012). The human milk microbiome changes over lactation and is shaped by maternal weight and mode of delivery. American Journal of Clinical Nutrition, 96(3), 544–551.
3. Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. Nature Methods, 13(7), 581–583.
4. Davis, N. M., Proctor, D. M., Holmes, S. P., Relman, D. A., & Callahan, B. J. (2018). Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. Microbiome, 6, 226.
5. Douglas, C. A., Ivey, K. L., Papanicolas, L. E., Best, K. P., Muhlhausler, B. S., & Rogers, G. B. (2020). DNA extraction approaches substantially influence the assessment of the human breast milk microbiome. Scientific Reports, 10, 123.
6. Fernandes, A. D., Reid, J. N., Macklaim, J. M., McMurrough, T. A., Edgell, D. R., & Gloor, G. B. (2014). Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. Microbiome, 2, 15.
7. Filatava, E. J., Liu, Z., Xie, J., Tran, D.-B., Chen, K., El Habbal, N., Weinstock, G., Zhou, Y., & Gregory, K. E. (2023). The preterm human milk microbiota fluctuates by postpartum week and is characterized by gestational age and maternal BMI. mBio, 14(6), e0210623.
8. Hunt, K. M., Foster, J. A., Forney, L. J., Schütte, U. M. E., Beck, D. L., Abdo, Z., Fox, L. K., Williams, J. E., McGuire, M. K., & McGuire, M. A. (2011). Characterization of the diversity and temporal stability of bacterial communities in human milk. PLoS ONE, 6(6), e21313.
9. Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology, 15, 550.
10. Lyons, K. E., O'Shea, C.-A., Grimaud, G., Ryan, C. A., Dempsey, E., Kelly, A. L., Ross, R. P., & Stanton, C. (2022). The human milk microbiome aligns with lactation stage and not birth mode. Scientific Reports, 12, 5598.
11. McMurdie, P. J., & Holmes, S. (2013). phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. PLoS ONE, 8(4), e61217.
12. Moossavi, S., & Azad, M. B. (2020). Origins of human milk microbiota: new evidence and arising questions. Gut Microbes, 12(1), 1667722.
13. Pang, W. W., & Hartmann, P. E. (2007). Initiation of human lactation: secretory differentiation and secretory activation. Journal of Mammary Gland Biology and Neoplasia, 12(4), 211–221.
14. Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., & Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Research, 41(D1), D590–D596.
15. Salter, S. J., Cox, M. J., Turek, E. M., Calus, S. T., Cookson, W. O., Moffatt, M. F., Turner, P., Parkhill, J., Loman, N. J., & Walker, A. W. (2014). Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biology, 12, 87.
16. Woortman, M. A., Sun, H., Wang, J., Godoy-Vitorino, F., Meléndez, A. J. R., Rivera, M. C., Piñero, E. E. A., Engelhardt, K., Kleinman, L. C., & Dominguez-Bello, M. G. (2025). Day/night fluctuations of breast milk bioactive factors and microbiome. Frontiers in Nutrition, 12, 1618784.