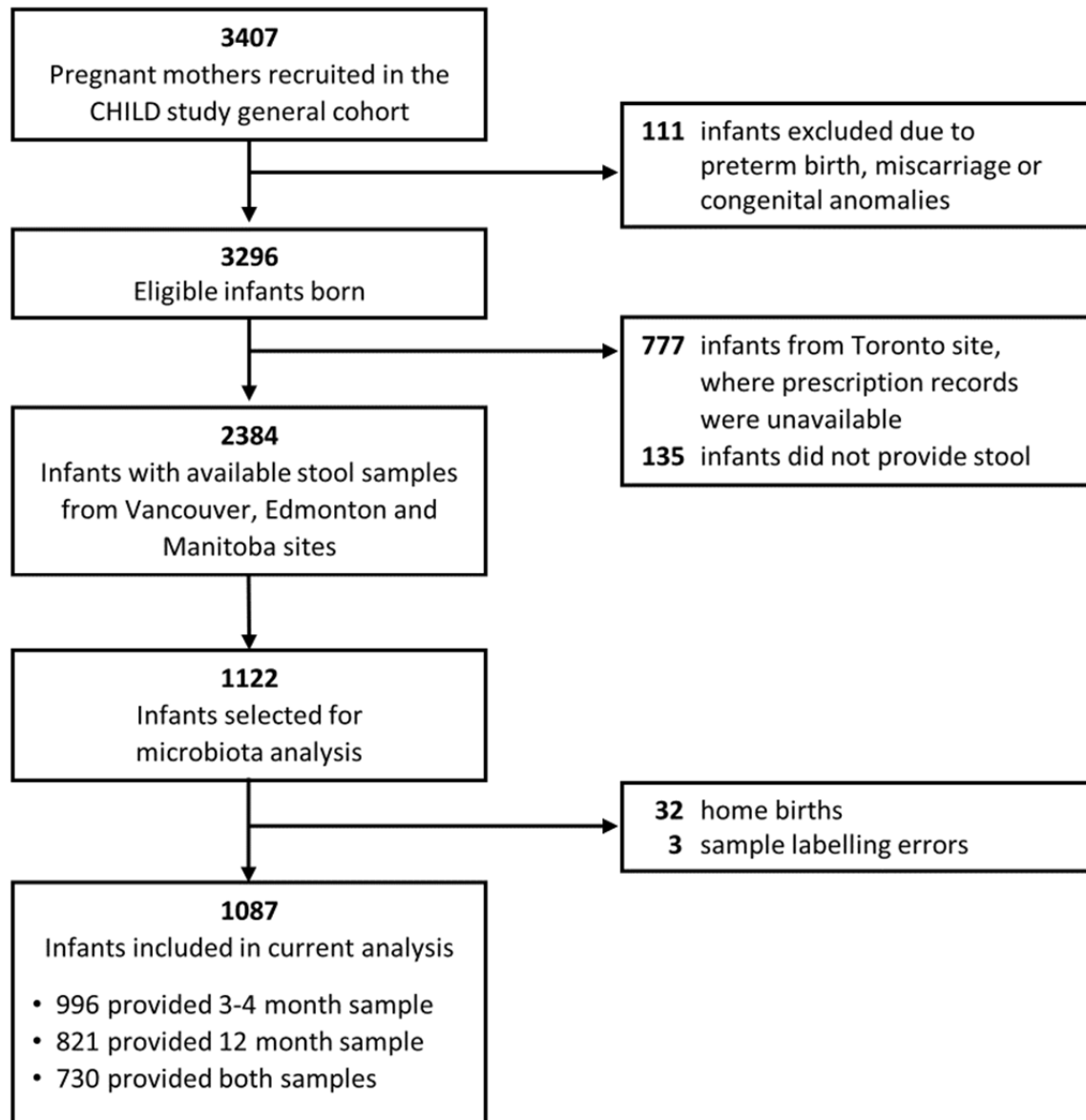# Supplementary Online Content

Forbes JD, Azad MB, Vehling L, et al; Canadian Healthy Infant Longitudinal Development (CHILD) Study investigators. Association of exposure to formula in the hospital and subsequent infant feeding practices with gut microbiota and risk of overweight in the first year of life. *JAMA Pediatr.* Published online June 4, 2018. doi:10.1001/jamapediatrics.2018.1161

This supplementary material has been provided by the authors to give readers additional information about their work.

**eFigure 1.** Flow Diagram Summarizing Selection of CHILD Study Infants Included in the Current Analysis



```
┌─────────────────────────────┐
│          3407               │
│  Pregnant mothers recruited │
│   in the CHILD study        │      ┌──────────────────────────────────┐
│   general cohort            │─────▶│ 111  infants excluded due to     │
└─────────────────────────────┘      │      preterm birth, miscarriage  │
              │                      │      or congenital anomalies     │
              ▼                      └──────────────────────────────────┘
┌─────────────────────────────┐
│          3296               │
│   Eligible infants born     │      ┌──────────────────────────────────┐
└─────────────────────────────┘      │ 777  infants from Toronto site,  │
              │                ─────▶│      where prescription records  │
              ▼                      │      were unavailable            │
┌─────────────────────────────┐      │ 135  infants did not provide     │
│          2384               │      │      stool                       │
│ Infants with available      │      └──────────────────────────────────┘
│ stool samples from          │
│ Vancouver, Edmonton and     │
│ Manitoba sites              │
└─────────────────────────────┘
              │
              ▼
┌─────────────────────────────┐
│          1122               │
│  Infants selected for       │      ┌──────────────────────────────────┐
│  microbiota analysis        │─────▶│ 32  home births                  │
└─────────────────────────────┘      │  3  sample labelling errors      │
              │                      └──────────────────────────────────┘
              ▼
┌─────────────────────────────┐
│          1087               │
│ Infants included in current │
│ analysis                    │
│ • 996 provided 3-4 month    │
│   sample                    │
│ • 821 provided 12 month     │
│   sample                    │
│ • 730 provided both samples │
└─────────────────────────────┘
```

**eFigure 2.** Microbial Community Structure of 3-Month and 12-Month Microbiota Based on Breastfeeding Status at 3-4 Months and Infant Diet at 6 months, Respectively, as Measured by Principal Components Analysis and tested by PERMANOVA.
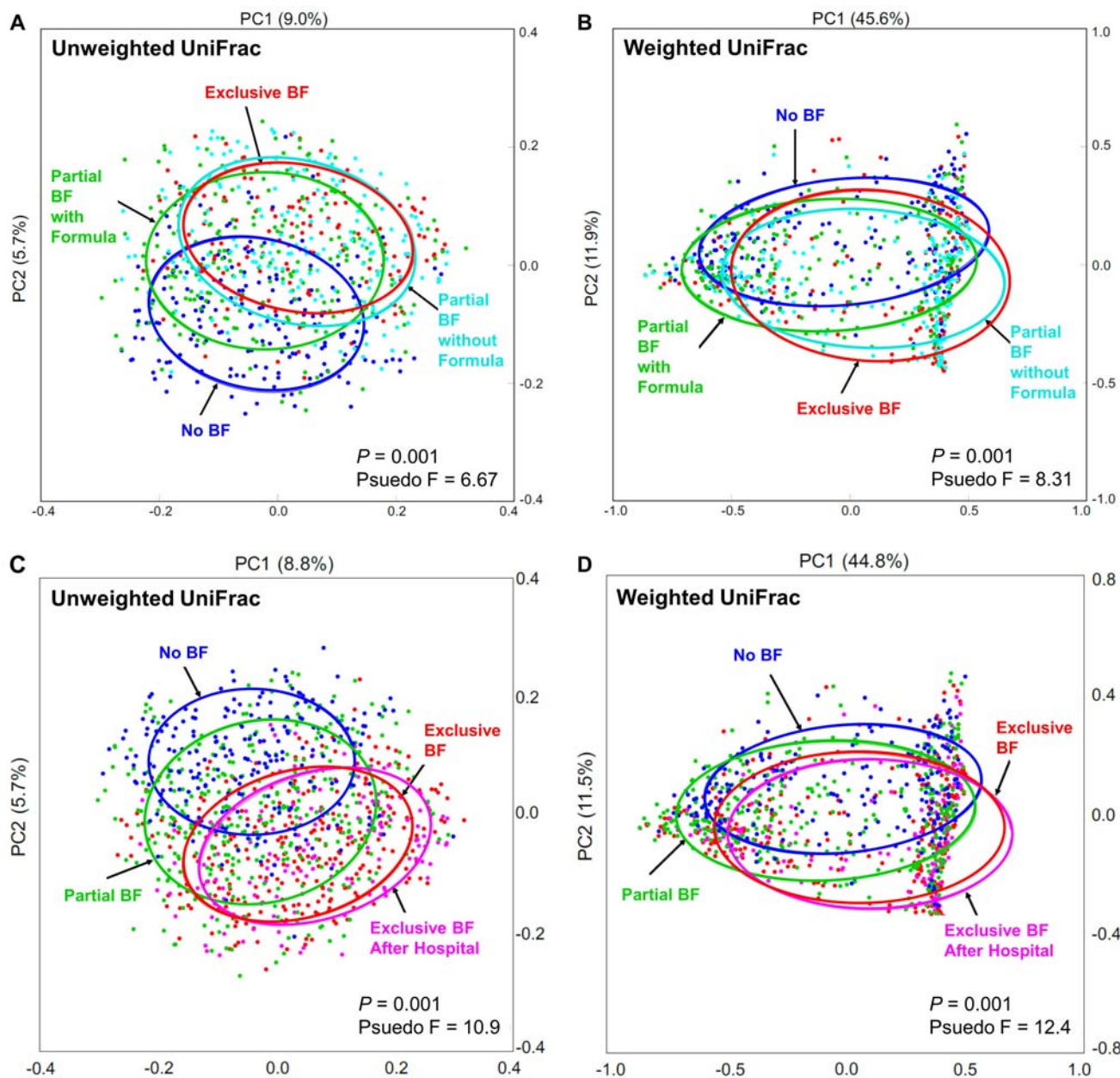


Principal coordinate analysis based on (A,C) unweighted or (B,D) weighted Unifrac distances, with community structure differences tested by PERMANOVA with 999 permutations.

# eFigure 3. Infant Gut Microbiota at 12 Months According to Breastfeeding (BF) Duration[#]
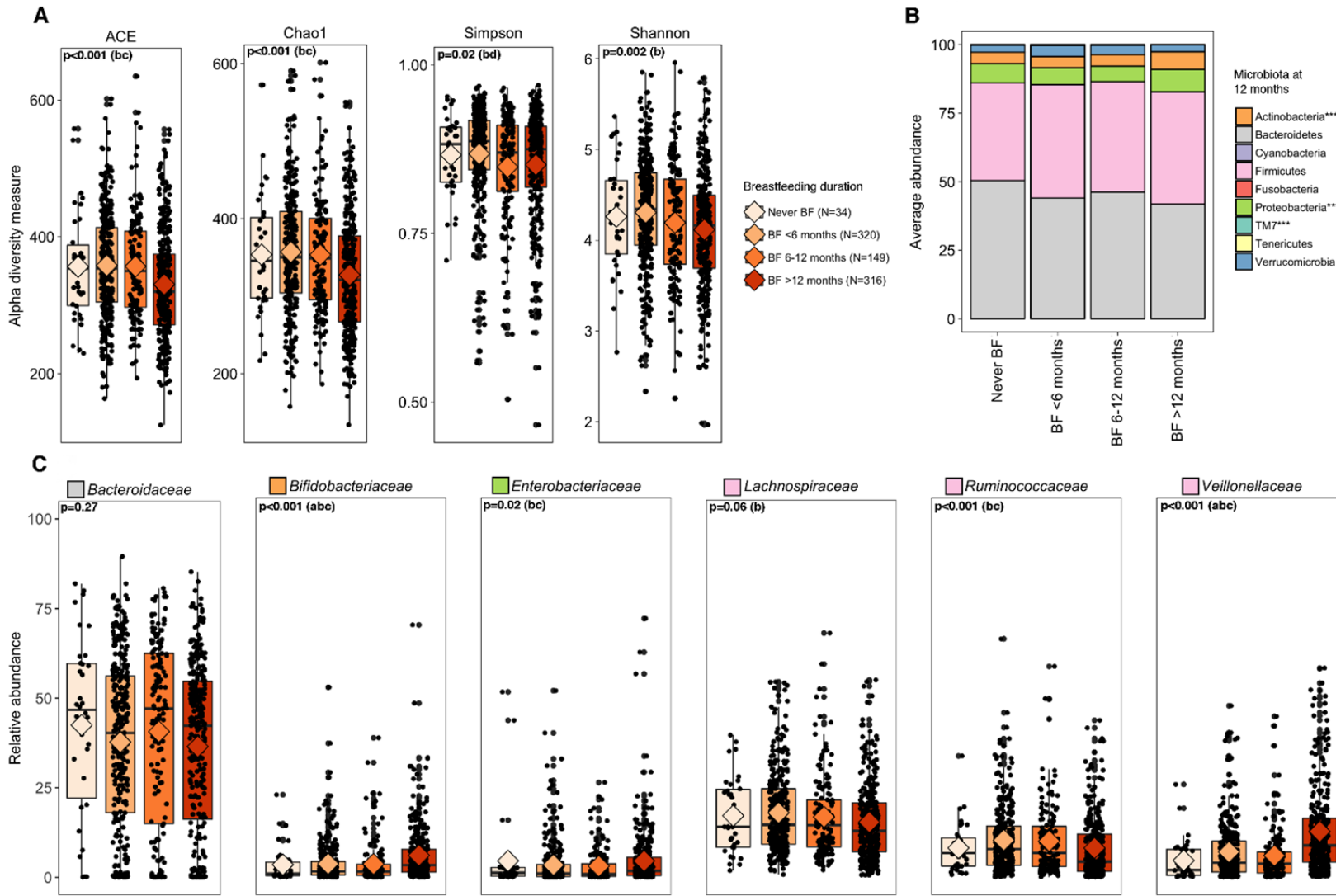


(A) Alpha diversity evaluated by richness (ACE, Chao1) and diversity (Simpson, Shannon). Median estimates compared across feeding groups using the Kruskal–Wallis test and Dunn's post-hoc tests for multiple comparisons. Boxes indicate interquartile range, lines indicate medians, diamonds indicate means and whiskers represent range. P-values represent overall FDR corrected P-values: *p<0.05, **p<0.01, ***p<0.001. (B) Mean phylum-level composition. (C). Relative abundance of dominant taxa across feeding groups. Significant pairwise comparisons for panels: [a]BF < 6 months/BF > 12 months; [b]BF 6 – 12 months/BF > 12 months; [c]Never BF/BF > 12 months; no significant differences were observed between BF <6 months/BF 6 – 12 months, Never BF/BF < 6 months or Never BF/BF 6 – 12 months. Breastfeeding refers to breast milk feeding, regardless of feeding mode (at the breast or from a bottle).

**eFigure 4.** Infant Gut Microbiota Characterization at 12 Months According to Infant Weight Status at 12 Months
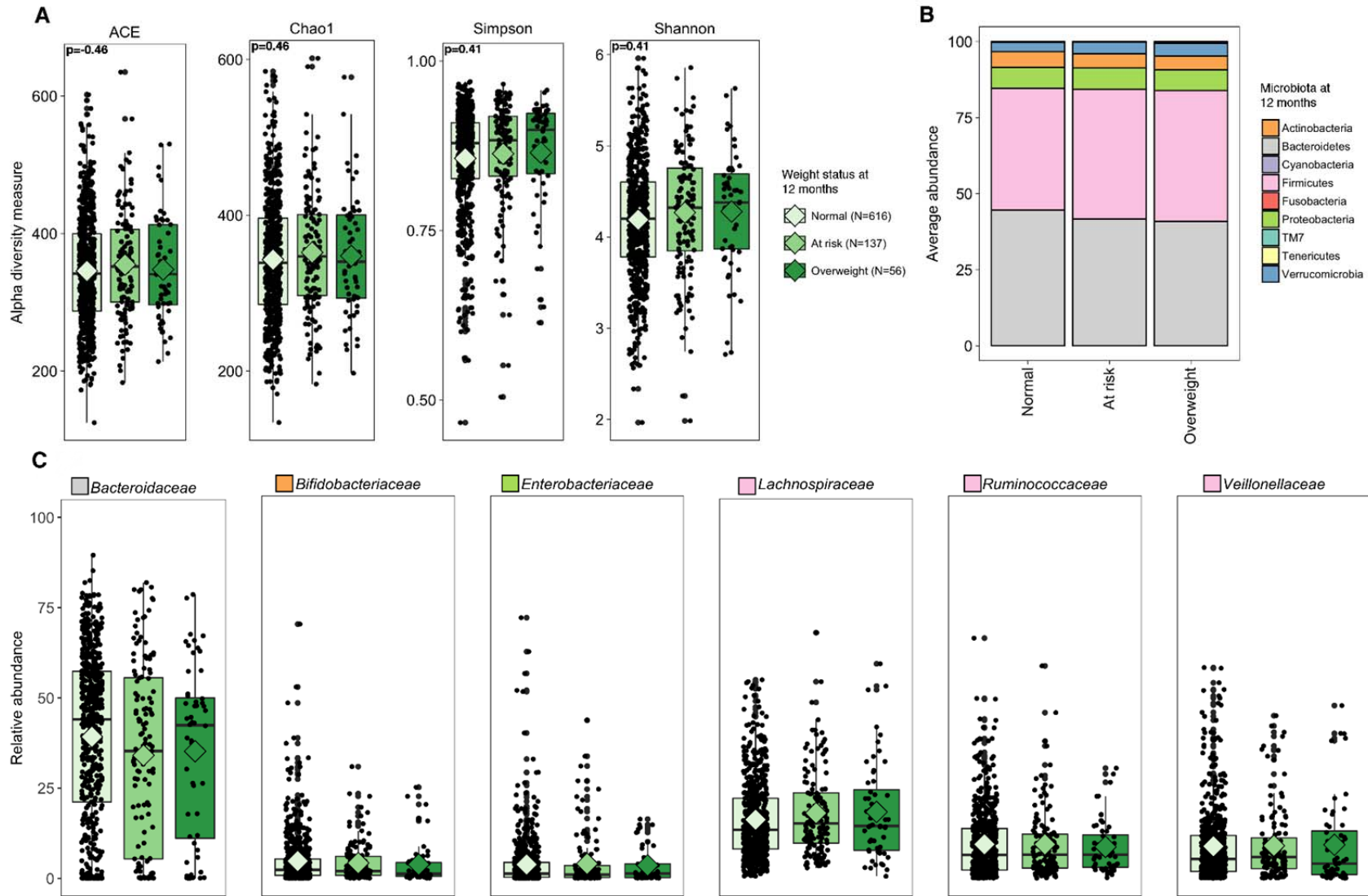


(A) Alpha diversity evaluated by richness (ACE, Chao1) and diversity (Simpson, Shannon). Median estimates compared across weight status using the Kruskal–Wallis test and Dunn's post-hoc tests for multiple comparisons. Boxes indicate interquartile range, lines indicate medians, diamonds indicate means and whiskers represent range. P-values represent overall FDR corrected P-values: *p<0.05, **p<0.01, ***p<0.001. (B) Mean phylum-level composition. (C). Relative abundance of dominant taxa across weight status groups. Significant pairwise comparisons: [a]Normal/At Risk; [b]Normal/Overweight; [c]At Risk/Overweight. Breastfeeding refers to breast milk feeding, regardless of feeding mode (at the breast or from a bottle).

**eFigure 5.** Association of Key Microbiota Measures at 3 and 12 Months With Infant Weight Status at 12 Months



OR (95%) for being Overweight or At Risk of Overweight (WFL >85th percentile) at 12 months
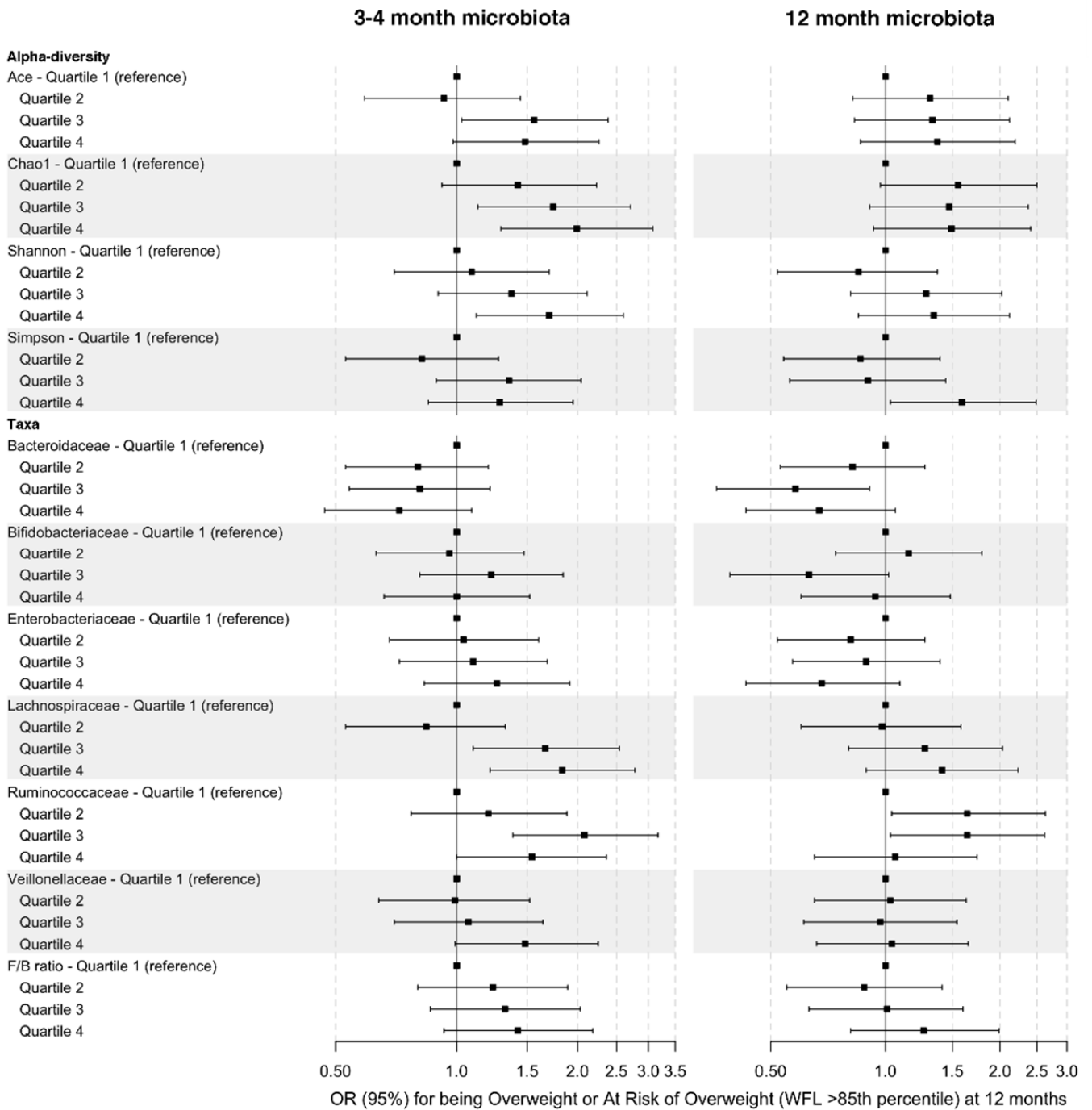
F/B ratio, Firmicutes/Bacteroidetes ratio; CI, confidence interval; OR, unadjusted odds ratio.

**eFigure 6.** Associations and Hypothesized Mechanisms Linking Infant Feeding Practices, Gut Microbiota and Obesity
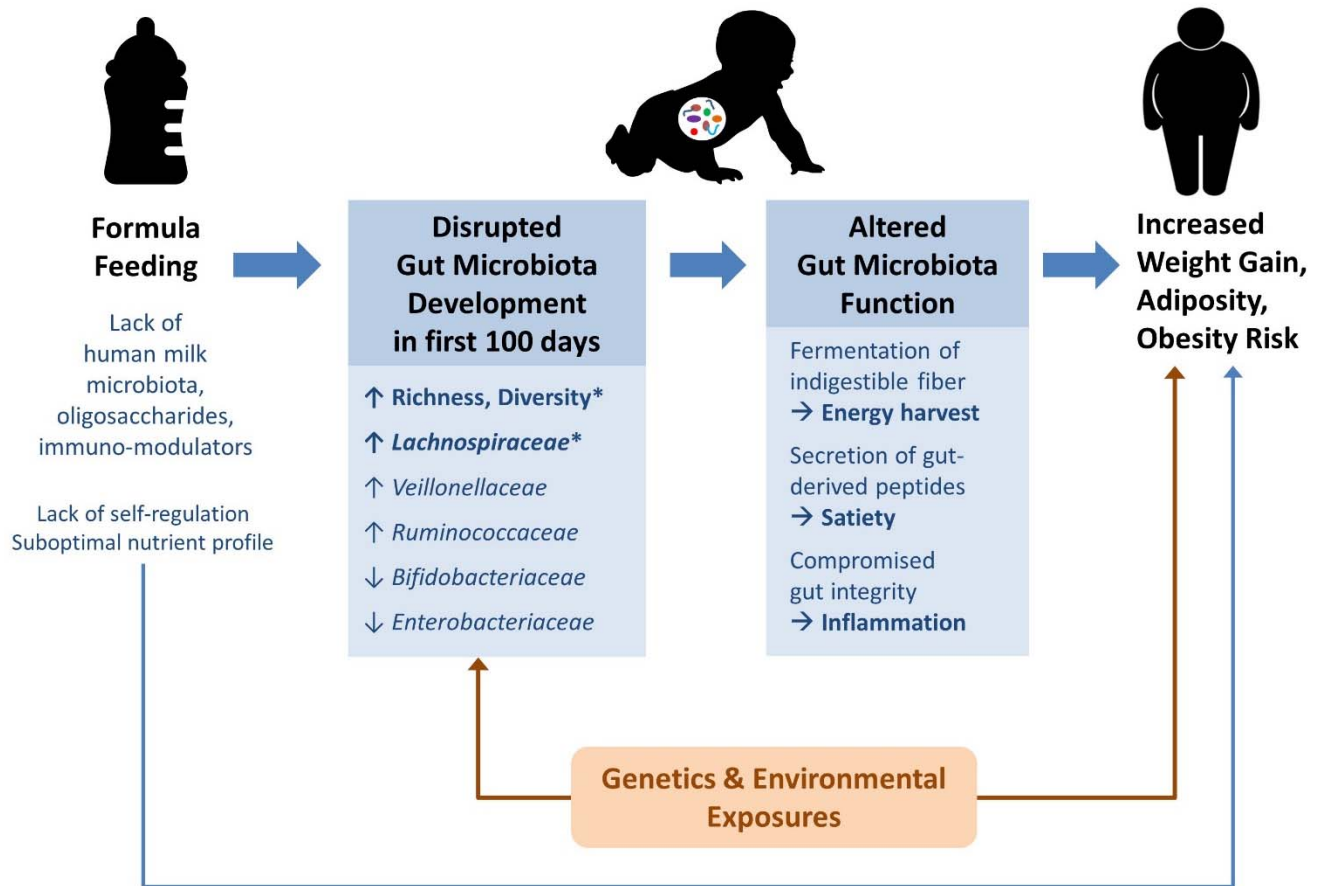


Results from this study show that formula-feeding is strongly associated with increased overweight risk at 1 year of age, and provide evidence that early changes in the gut microbiota (i.e. during the first 3-4 months of life) contribute to this association. Infant formulas differ from human milk in composition and bioactivity, causing changes in gut microbial communities that likely lead to altered metabolic networks affecting energy harvest, satiety and inflammation. These physiological changes, along with host genotype and environmental exposures, influence infant weight gain, adiposity and obesity risk later in life. Aside from these microbiota-related pathways, formula may influence weight gain through other mechanisms related to its composition (eg. higher protein content compared to human milk) or delivery to the infant (e.g. bottle feeding may discourage self-regulation compared to suckling at the breast). *Indicates microbiota features significantly associated with both formula feeding and risk of overweight in this study.

**eTable 1.** Characteristics of Participants Included in the Current Study and the General CHILD Cohort

| | Subset for current study (N=1087) | | General CHILD cohort (N=3296) | |
|---|---|---|---|---|
| **Breastfeeding duration** | 10.26 | [6.81] | 10.39 | [6.81] |
| **Exclusive breastfeeding duration** | 2.90 | [2.37] | 3.16 | [2.31] |
| **Weight for length (WFL) z-score** | 0.29 | [1.08] | 0.25 | [1.04] |
| **Maternal race** | | | | |
| Asian | 164 | (15.2) | 508 | (15.7) |
| Caucasian | 817 | (75.8) | 2359 | (72.9) |
| FN | 48 | (4.5) | 143 | (4.4) |
| Other | 49 | (4.5) | 225 | (7.0) |
| **Maternal postsecondary degree** | | | | |
| No | 230 | (21.8) | 746 | (23.7) |
| Yes | 823 | (78.2) | 2407 | (76.3) |
| **Maternal smoking in pregnancy** | | | | |
| No | 967 | (91.5) | 2897 | (90.9) |
| Yes | 90 | (8.5) | 290 | (9.1) |
| **Maternal Healthy Eating Index** | | | | |
| < 70 | 306 | (30.4) | 990 | (33.1) |
| 70 to 75 | 240 | (23.9) | 682 | (22.8) |
| > 75 | 459 | (45.7) | 1323 | (44.2) |
| **Maternal weight class** | | | | |
| Underweight | 31 | (3.0) | 103 | (3.5) |
| Normal | 586 | (57.2) | 1760 | (59.9) |
| Overweight | 230 | (22.4) | 633 | (21.6) |
| Obese | 178 | (14.4) | 440 | (15.0) |
| **Dog in home** | | | | |
| No | 687 | (66.9) | 2142 | (69.7) |
| Yes | 340 | (33.1) | 930 | (30.3) |
| **Cat in home** | | | | |
| No | 745 | (72.6) | 2308 | (75.2) |
| Yes | 281 | (27.4) | 762 | (24.8) |
| **Older siblings** | | | | |
| No | 569 | (52.3) | 1772 | (53.9) |
| Yes | 518 | (47.7) | 1519 | (46.2) |
| **Infant sex** | | | | |
| Male | 580 | (53.4) | 1726 | (52.7) |
| Female | 507 | (46.6) | 1550 | (47.3) |
| **Birth weight (g)** | | | | |
| < 3000 | 162 | (15.2) | 527 | (16.5) |
| 3000 to < 3500 | 419 | (39.4) | 1243 | (38.9) |
| 3500 to < 4000 | 350 | (32.9) | 1024 | (32.0) |
| 4000 + | 133 | (12.5) | 403 | (12.6) |
| **Birth mode** | | | | |
| CS-Elective | 121 | (11.4) | 346 | (10.8) |
| CS-Emergency | 153 | (14.4) | 466 | (14.6) |
| Vaginal, IAP | 237 | (22.3) | 688 | (21.5) |
| Vaginal, no IAP | 553 | (52.0) | 1695 | (53.1) |
| **Oral antibiotics before 12 months** | | | | |
| No | 815 | (79.7) | 2266 | (80.3) |
| Yes | `208 | (20.3) | 557 | (19.7) |

BMI, body mass index; WFL, weight-for-length; FN, First Nations; CS, caesarean section; IAP, intrapartum antibiotic prophylaxis.
Values are n (%) or mean [standard deviation]. Percentages reflect proportion of non-missing data for each variable. Breastfeeding refers to breast milk feeding, regardless of feeding mode (at the breast or from a bottle).

**eTable 2.** Infant Feeding and Weight Variables Among Participants in the Subcohort (N* = 1087)

| Feeding exposure variables | | |
|---|---|---|
| **Breastfeeding at 3 months** | | |
| None (formula only) | 175 | (16.1) |
| Partial (breast milk + formula) | 323 | (29.7) |
| Exclusive after hospital (brief formula supplementation in hospital) | 179 | (16.5) |
| Exclusive (breast milk only) | 400 | (36.8) |
| *Missing* | 10 | (0.9) |
| **Breastfeeding status at time of 3-4 month sampling** | | |
| None (formula only) | 225 | (20.7) |
| Partial (breast milk + formula) | 367 | (33.8) |
| Exclusive after hospital (brief formula supplementation in hospital) | 150 | (13.8) |
| Exclusive (breast milk only) | 324 | (29.8) |
| *Missing* | 21 | (1.9) |
| **Breastfeeding at 6 months** | | |
| None (formula +/- food) | 264 | (24.3) |
| Partial breastfeeding with formula (breast milk + formula +/- food) | 307 | (28.2) |
| Partial breastfeeding without formula (breast milk + food) | 286 | (26.3) |
| Exclusive (breast milk only) | 183 | (16.8) |
| *Missing* | 47 | (4.3) |
| **Breastfeeding at 12 months** | | |
| No | 591 | (54.4) |
| Yes | 459 | (42.2) |
| *Missing* | 37 | (3.4) |
| **Breastfeeding duration** | | |
| Never breastfed | 49 | (4.5) |
| Breastfed < 6 months | 324 | (29.8) |
| Breastfed 6 – 12 months | 249 | (22.9) |
| Breastfed > 12 months | 459 | (42.2) |
| *Missing* | 6 | (0.05) |
| **Breastfeeding duration at time of 12 month sampling** | | |
| Never breastfed | 43 | (4.0) |
| Breastfed < 6 months | 200 | (18.4) |
| Breastfed > 6 months, but not currently breastfeeding | 365 | (33.6) |
| Currently breastfeeding | 367 | (33.8) |
| *Missing* | 112 | (10.3) |
| **Breastfeeding duration** (months; N = 1081) | 10.26 | [6.81] |
| **Exclusive breastfeeding duration** (months; N = 1058) | 2.90 | [2.37] |
| **Weight outcome variables** | | |
| **Weight class at 1 year** | | |
| Normal (WFL z-score < 85th percentile) | 778 | (71.6) |
| At risk (85th – 97th percentile) | 178 | (16.4) |
| Overweight (> 97th percentile) | 71 | (6.5) |
| *Missing* | 60 | (5.5) |
| **WFL z-score at 12 months** (N = 1027) | 0.29 | [1.08] |

WFL, weight-for-length. *N = number of infants with 3 month and/or 12 month microbiota data. Values are n (%) or mean [± standard deviation]. Breastfeeding refers to breast milk feeding, regardless of feeding mode (at the breast or from a bottle).

**eTable 3.** Prevalence of Potential Confounders and Associations With Breastfeeding and Overweight Risk

| | Overall Prevalence | | Exclusive Breastfeeding at 3 months | | Any Breastfeeding at 12 months | | At Risk or Overweight at 12 months: WFL >85th percentile | |
|---|---|---|---|---|---|---|---|---|
| | n/N | (%) | % | P Value | % | P Value | % | P Value |
| **Maternal race** | | | | | | | | |
| Asian | 164/1078 | (15.2) | 53.7 | | 51.6 | | 17.6 | |
| Caucasian | 817/1078 | (75.8) | 55.0 | 0.13 | 43.1 | 0.10 | 24.4 | 0.007 |
| FN | 48/1078 | (4.5) | 45.8 | | 36.2 | | 43.2 | |
| Other | 49/1078 | (4.5) | 39.6 | | 35.7 | | 24.4 | |
| **Maternal postsecondary degree** | | | | | | | | |
| No | 230/1053 | (21.8) | 45.9 | 0.008 | 30.3 | <0.001 | 31.3 | 0.008 |
| Yes | 823/1053 | (78.2) | 56.0 | | 47.7 | | 22.3 | |
| **Maternal smoking in pregnancy** | | | | | | | | |
| No | 967/1057 | (91.5) | 55.9 | <0.001 | 46.1 | <0.001 | 23.7 | 0.17 |
| Yes | 90/1057 | (8.5) | 30.3 | | 20.9 | | 31.3 | |
| **Maternal Healthy Eating Index 2010 Score** | | | | | | | | |
| < 70 | 306/1005 | (30.4) | 42.8 | | 33.2 | | 23.7 | |
| 70 to 75 | 240/1005 | (23.9) | 55.8 | <0.001 | 42.0 | <0.001 | 25.3 | 0.87 |
| > 75 | 459/1005 | (45.7) | 60.3 | | 52.1 | | 23.6 | |
| **Maternal weight class** | | | | | | | | |
| Underweight | 31/1025 | (3.0) | 54.8 | | 48.3 | | 20.7 | |
| Normal | 586/1025 | (57.2) | 59.2 | <0.001 | 49.4 | <0.001 | 21.7 | 0.05* |
| Overweight | 230/1025 | (22.4) | 55.0 | | 47.7 | | 24.4 | |
| Obese | 178/1025 | (17.4) | 34.3 | | 24.7 | | 28.9 | |
| **Dog in home** | | | | | | | | |
| No | 687/1027 | (66.9) | 57.2 | 0.003 | 47.5 | 0.002 | 24.0 | 0.94 |
| Yes | 340/1027 | (33.1) | 47.4 | | 37.0 | | 23.6 | |
| **Cat in home** | | | | | | | | |
| No | 745/1026 | (72.6) | 54.4 | 0.67 | 44.2 | 0.93 | 23.5 | 0.68 |
| Yes | 281/1026 | (27.4) | 52.7 | | 43.6 | | 25.0 | |
| **Older siblings** | | | | | | | | |
| No | 569/1087 | (52.3) | 51.9 | 0.27 | 44.5 | 0.56 | 25.4 | 0.40 |
| Yes | 518/1087 | (47.7) | 55.4 | | 42.7 | | 23.0 | |
| **Infant sex** | | | | | | | | |
| Male | 507/1087 | (46.6) | 57.1 | 0.04 | 43.8 | 1 | 23.8 | 0.83 |
| Female | 580/1087 | (53.4) | 50.6 | | 43.6 | | 24.6 | |
| **Birth weight (g)** | | | | | | | | |
| < 3000 | 162/1064 | (15.2) | 51.3 | | 43.3 | | 13.8 | |
| 3000 to < 3500 | 419/1064 | (39.4) | 54.1 | 0.91 | 41.5 | 0.79 | 17.7 | <0.001 |
| 3500 to < 4000 | 350/1064 | (32.9) | 54.0 | | 45.1 | | 30.5 | |
| 4000 + | 133/1064 | (12.5) | 51.9 | | 44.6 | | 40.6 | |
| **Birth mode** | | | | | | | | |
| CS-Elective | 121/1064 | (11.4) | 45.5 | | 41.2 | | 25.0 | |
| CS-Emergency | 153/1064 | (14.4) | 54.9 | 0.31 | 43.0 | 0.95 | 27.5 | 0.16 |
| Vaginal, IAP | 237/1064 | (22.3) | 54.7 | | 44.2 | | 28.0 | |
| Vaginal, no IAP | 553/1064 | (52.0) | 54.4 | | 43.9 | | 21.2 | |
| **Any oral antibiotics between birth and 12 months** | | | | | | | | |
| No | 815/1023 | (79.7) | 53.6 | 0.55 | 45.9 | 0.05 | 23.8 | 0.63 |
| Yes | `208/1023 | (20.3) | 51.0 | | 37.9 | | 25.7 | |

FN, First Nations; CS, caesarean section; IAP, intrapartum antibiotic prophylaxis, WFL, weight-for-length. Comparisons by chi-squared test or *Cochran-Armitage test for trend. Breastfeeding refers to breast milk feeding, regardless of feeding mode (at the breast or from a bottle).

**eTable 4.** Sensitivity Analyses: Association of Infant Feeding Practices With Infant Weight Status at 12 Months

| | Model 1[1]<br>OR (95%CI) | Model 2[2]<br>OR (95%CI) | Model 3[3]<br>OR (95%CI) | Model 4[4]<br>OR (95%CI) | OR (95%CI) | OR (95%CI) | Beta (95%CI) |
|---|---|---|---|---|---|---|---|
| **Outcome** | WFL >85th percentile | WFL >85th percentile | WFL >85th percentile | WFL >85th percentile | WFL >85th percentile | WFL >85th percentile | WFL z-score |
| **Sensitivity Analysis** | Maternal BMI, Site | Other Maternal Factors: Maternal Smoking, Education, Ethnicity, Site | Microbiota Factors: Caesarean section, Dog, Infant Sex, Oral Antibiotics, Site | Maternal BMI, Other Maternal Factors, Microbiota Factors | Model 4 + Further Adjustment for Birth Weight | Model 4 Excluding Never-Breastfed Infants | Model 4 with Alternative Outcome: WFL z-score |
| **Breastfeeding at 3 months** | N = 990 | N = 985 | N = 913 | N = 879 | N=868 | N=814 | N = 879 |
| None | 2.04 (1.31 –3.19) | 2.00 (1.27 – 3.12) | 2.34 (1.49 –3.68) | 2.04 (1.25 – 3.32) | 2.15 (1.30 – 3.56) | 2.33 (1.29 – 4.16) | 0.30 (0.08 – 0.51) |
| Partial | 1.55 (1.07 – 2.25) | 1.53 (1.06 – 2.21) | 1.67 (1.13 –2.45) | 1.63 (1.10 – 2.45) | 1.77 (1.17 – 2.69) | 1.64 (1.09 – 2.47) | 0.26 (0.09 – 0.43) |
| Exclusive after hospital | 1.06 (0.66 – 1.69) | 1.07 (0.67 – 1.70) | 1.17 (0.71 – 1.91) | 1.13 (0.67 – 1.87) | 1.12 (0.65 – 1.89) | 1.14 (0.68 – 1.89) | -0.04 (-0.25 – 0.16) |
| Exclusive | 1.00 (reference) | 1.00 (reference) | 1.00 (reference) | 1.00 (reference) | 1.00 (reference) | 1.00 (reference) | 0.00 (reference) |
| | | | | | | | |
| **Breastfeeding at 6 months** | N = 973 | N = 967 | N = 901 | N = 868 | N = 856 | N = 803 | N = 868 |
| None | 1.94 (1.20 – 3.21) | 1.92 (1.18 – 3.17) | 1.94 (1.18 – 3.26) | 1.65 (0.98 – 2.82) | 1.57 (0.92 – 2.73) | 1.68 (0.97 – 2.96) | 0.28 (0.05 – 0.51) |
| Partial with formula | 1.63 (1.02 – 2.65) | 1.65 (1.03 – 2.68) | 1.65 (1.03 – 2.74) | 1.43 (0.87 – 2.39) | 1.46 (0.88 – 2.48) | 1.44 (0.88 – 2.41) | 0.19 (-0.02 – 0.40) |
| Partial without formula | 1.11 (0.68 – 1.85) | 1.16 (0.71 – 1.92) | 1.06 (0.64 – 1.80) | 0.95 (0.56 – 1.63) | 0.82 (0.47 – 1.43) | 0.97 (0.57 – 1.66) | 0.02 (-0.20 – 0.23) |
| Exclusive | 1.00 (reference) | 1.00 (reference) | 1.00 (reference) | 1.00 (reference) | 1.00 (reference) | 1.00 (reference) | 0.00 (reference) |
| | | | | | | | |
| **Breastfeeding duration** | N = 950 | N = 945 | N = 876 | N = 844 | N = 832 | N/A | N = 844 |
| < 6 months* | 1.89 (1.27 – 2.81) | 1.77 (1.19 – 2.63) | 1.84 (1.22 –2.77) | 1.64 (1.06 – 2.52) | 1.70 (1.08 – 2.66) | N/A | 0.27 (0.08 – 0.45) |
| 6 to < 12 months | 1.60 (1.11 – 2.30) | 1.59 (1.11 – 2.29) | 1.63 (1.12 – 2.37) | 1.47 (0.99 – 1.28) | 1.57 (1.04 – 2.36) | N/A | 0.19 (0.03 – 0.36) |
| ≥ 12 months | 1.00 (reference) | 1.00 (reference) | 1.00 (reference) | 1.00 (reference) | 1.00 (reference) | N/A | 0.00 (reference) |

OR, odds ratio; CI, confidence interval; WFL, weight for length. [1]Adjusted for maternal BMI and study site. [2]Adjusted for maternal smoking, post-secondary education and ethnicity and site. [3]Adjusted for caesarean section, dog in household, infant sex, any oral antibiotics between 0 and 12 months and study site. [4]Adjusted for maternal BMI, smoking, post-secondary education and ethnicity and site, caesarean section, dog in household, infant sex and any oral antibiotics between 0 and 12 months. *Excludes infants who were never breastfed. Breastfeeding refers to breast milk feeding, regardless of feeding mode (at the breast or from a bottle).

**eTable 5.** Median Relative Abundance of Abundant† Taxa in Gut Microbiota at 3-4 Months According to Feeding Status[#]

| Phylum / *Family* | (A) No Breastfeeding N=222 | (B) Partial Breastfeeding N=340 | (C) Exclusive Breastfeeding after Hospital N=137 | (D) Exclusive Breastfeeding N=291 | Overall pFDR | A vs B | A vs C | A vs D | B vs C | B vs D | C vs D |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Actinobacteria** | 3.829 | 5.468 | 5.368 | 8.456 | *** | a |  | c |  |  | f |
| *Actinomycetaceae* | 0.039 | 0.031 | 0.016 | 0.016 | ** |  | b | c | d | e |  |
| *Bifidobacteriaceae* | 3.006 | 4.882 | 4.342 | 8.315 | *** | a | b | c |  | e | f |
| *Coriobacteriaceae* | 0.079 | 0.054 | 0.023 | 0.016 | *** | a | b | c | d | e |  |
| *Micrococcaceae* | 0.008 | 0.008 | 0.008 | 0.016 | *** |  |  | c | d | e |  |
| **Bacteroidetes** | 22.446 | 38.887 | 0.908 | 4.061 | *** | a | b |  | d | e |  |
| *Bacteroidaceae* | 14.208 | 26.689 | 0.511 | 2.686 | ** | a |  |  | d | e |  |
| *Porphyromonadaceae* | 0.008 | 0.008 | 0.000 | 0.000 | *** |  | b | c | d | e |  |
| *Prevotellaceae* | 0.000 | 0.000 | 0.000 | 0.000 | * |  |  | c |  |  | f |
| *Rikenellaceae* | 0.000 | 0.000 | 0.000 | 0.000 | ** |  | b | c | d | e |  |
| **Firmicutes** | 31.996 | 19.703 | 18.073 | 20.405 | *** | a | b | c |  |  |  |
| Clostridiales (other) | 0.000 | 0.000 | 0.000 | 0.000 | *** | a | b | c | d | e |  |
| Clostridiales (unclassified) | 0.000 | 0.000 | 0.000 | 0.000 | *** | a | b | c | d | e |  |
| *Clostridiaceae* | 0.785 | 0.294 | 0.139 | 0.287 | *** | a | b | c |  |  |  |
| *Enterococcaceae* | 0.023 | 0.023 | 0.023 | 0.016 |  |  |  |  |  |  |  |
| *Erysipelotrichaceae* | 0.215 | 0.023 | 0.000 | 0.008 | *** | a | b | c | d | e |  |
| *Gemellaceae* | 0.008 | 0.000 | 0.008 | 0.000 | ** | a |  |  | d |  |  |
| *Lachnospiraceae* | 7.942 | 3.022 | 0.349 | 0.395 | *** | a | b | c | d | e |  |
| Lactobacillales (other) | 0.016 | 0.000 | 0.000 | 0.000 | *** | a | b |  | d | e |  |
| *Ruminococcaceae* | 1.961 | 0.334 | 0.008 | 0.008 | *** | a | b | c | d | e |  |
| *Streptococcaceae* | 0.587 | 0.541 | 0.575 | 0.619 |  |  |  |  |  |  |  |
| *Veillonellaceae* | 7.899 | 4.531 | 2.610 | 3.047 | *** | a | b | c | d | e |  |
| **Proteobacteria** | 11.271 | 15.922 | 32.261 | 26.492 | *** | a | b | c | d | e |  |
| *Alcaligenaceae* | 0.000 | 0.000 | 0.000 | 0.000 |  |  |  |  |  |  |  |
| *Enterobacteriaceae* | 9.049 | 13.904 | 29.801 | 24.462 | *** | a | b | c | d | e | f |
| *Pasteurellaceae* | 0.008 | 0.016 | 0.086 | 0.101 | *** | a | b | c | d | e |  |
| **Verrucomicrobia** | 0.000 | 0.000 | 0.000 | 0.000 | *** | a | b | c | d | e |  |
| *Verrucomicrobiaceae* | 0.000 | 0.000 | 0.000 | 0.000 | *** | a | b | c | d | e |  |
| Firmicutes/ Bacteroidetes ratio | 1.54 | 0.56 | 17.63 | 5.53 | *** | a |  |  | d | e |  |

FDR, false discovery rate. †Taxa with median abundance >0% in 3 month and/or 12 month microbiota. [#]Feeding status at the time of sample collection. Overall comparisons by rank-based nonparametric Kruskal–Wallis test with FDR correction for multiple comparisons; *p<0.05, **p<0.01, ***p<0.001. Significant differences are possible when all medians are 0.000 because this is a rank-based test. Pairwise comparisons by Dunn's post-hoc tests for multiple comparisons (significant pairwise differences): [a]No breastfeeding (BF)/Partial BF; [b]No BF/Exclusive BF after hospital; [c]No BF/Exclusive BF; [d]Partial BF/Exclusive BF after hospital; [e]Partial BF/Exclusive BF; [f]Exclusive BF after hospital/Exclusive BF. Breastfeeding refers to breast milk feeding, regardless of feeding mode (at the breast or from a bottle).

**eTable 6.** Pairwise PERMANOVA Analyses of Infant Microbiota According to Feeding Status at 3-4 Months and 6 Months

| | Unweighted Unifrac | | Weighted Unifrac | |
|---|---|---|---|---|
| | Pseudo-F | *P* Value | Pseudo-F | *P* Value |
| **BF at 3-4 months** | **Microbiota at 3-4 months** | | | |
| No BF vs Partial BF | 8.11 | 0.001 | 8.46 | 0.001 |
| No BF vs Exclusive BF after hospital | 18.46 | 0.001 | 16.99 | 0.001 |
| No BF vs Exclusive BF | 22.54 | 0.001 | 18.63 | 0.001 |
| Exclusive BF after hospital vs Partial BF | 8.15 | 0.001 | 14.56 | 0.001 |
| Partial BF vs Exclusive BF | 8.16 | 0.001 | 14.03 | 0.001 |
| Exclusive BF after hospital vs Exclusive BF | 1.42 | 0.07 | 0.24 | 0.24 |
| **Diet at 6 months** | **Microbiota at 12 months** | | | |
| No BF vs Partial BF with formula | 6.67 | 0.001 | 5.74 | 0.002 |
| No BF vs Partial BF without formula | 13.37 | 0.001 | 12.74 | 0.001 |
| No BF vs Exclusive BF | 12.03 | 0.001 | 10.95 | 0.001 |
| Partial BF with formula vs Partial BF without formula | 3.76 | 0.001 | 9.72 | 0.001 |
| Partial BF with formula vs Exclusive BF | 3.43 | 0.001 | 9.81 | 0.001 |
| Partial BF without formula vs Exclusive BF | 0.93 | 0.59 | 0.40 | 0.78 |

BF, breastfeeding. Pairwise community structure differences tested by PERMANOVA with 999 permutations based on unweighted or weighted Unifrac distances.

**eTable 7.** Median Relative Abundance of Abundant† Taxa in Fecal Microbiota of Infants at 12 Months According to Feeding Status at 6 Months

| Phylum<br>*Family* | (A)<br>No Breastfeeding<br>N=190 | (B)<br>Partial Breastfeeding with Formula<br>N=248 | (C)<br>Partial Breastfeeding without Formula<br>N=218 | (D)<br>Exclusive Breastfeeding<br>N=147 | Overall pFDR | Pairwise pFDR A vs B | A vs C | A vs D | B vs C | B vs D | C vs D |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Actinobacteria** | 1.724 | 2.134 | 3.212 | 3.293 | ** | | b | c | d | e | |
| *Actinomycetaceae* | 0.008 | 0.008 | 0.008 | 0.008 | | | | | | | |
| *Bifidobacteriaceae* | 1.422 | 1.809 | 2.896 | 2.796 | ** | | b | c | d | e | |
| *Coriobacteriaceae* | 0.117 | 0.098 | 0.085 | 0.085 | | | | | | | |
| *Micrococcaceae* | 0.000 | 0.000 | 0.000 | 0.000 | | | | | | | |
| **Bacteroidetes** | 49.853 | 52.703 | 45.714 | 45.054 | ** | | b | c | d | e | |
| *Bacteroidaceae* | 42.131 | 46.911 | 39.098 | 41.060 | | | | | d | e | |
| *Porphyromonadaceae* | 0.125 | 0.082 | 0.016 | 0.008 | | | b | c | d | e | |
| *Prevotellaceae* | 0.008 | 0.008 | 0.012 | 0.015 | | | | | | | |
| *Rikenellaceae* | 0.008 | 0.008 | 0.004 | 0.008 | | | | | | | |
| **Firmicutes** | 36.420 | 34.671 | 39.050 | 36.364 | | | | | d | | |
| Clostridiales (other) | 0.008 | 0.008 | 0.008 | 0.008 | | | b | | | | |
| Clostridiales (unclassified) | 0.808 | 0.998 | 0.603 | 0.521 | | | | | | | |
| *Clostridiaceae* | 0.387 | 0.515 | 0.560 | 0.544 | | | | | | | |
| *Enterococcaceae* | 0.000 | 0.000 | 0.004 | 0.000 | | | | | | | |
| *Erysipelotrichaceae* | 0.526 | 0.452 | 0.420 | 0.350 | | | | c | | | |
| *Gemellaceae* | 0.008 | 0.008 | 0.008 | 0.008 | | | | | | | |
| *Lachnospiraceae* | 14.531 | 13.372 | 14.275 | 12.894 | | | | | | | |
| Lactobacillales (other) | 0.008 | 0.015 | 0.008 | 0.008 | | | | | | | |
| *Ruminococcaceae* | 7.275 | 6.433 | 5.225 | 5.713 | | | | | | | |
| *Streptococcaceae* | 0.257 | 0.285 | 0.337 | 0.312 | | | | | | | |
| *Veillonellaceae* | 3.040 | 4.611 | 7.644 | 7.078 | *** | a | b | c | d | e | |
| **Proteobacteria** | 3.611 | 3.782 | 5.104 | 5.706 | ** | | b | c | d | e | |
| *Alcaligenaceae* | 0.266 | 0.719 | 1.057 | 1.188 | | | b | c | | | |
| *Enterobacteriaceae* | 1.167 | 0.928 | 1.446 | 1.391 | | | | | d | | |
| *Pasteurellaceae* | 0.105 | 0.194 | 0.279 | 0.249 | * | a | b | c | | | |
| **Verrucomicrobia** | 0.023 | 0.016 | 0.015 | 0.016 | | | | | | | |
| *Verrucomicrobiaceae* | 0.023 | 0.016 | 0.015 | 0.016 | | | | | | | |
| Firmicutes/ Bacteroidetes ratio | 0.71 | 0.68 | 0.87 | 0.85 | ** | | b | | d | e | |

FDR, false discovery rate. †Taxa with median abundance >0% in 3 month and/or 12 month microbiota. Overall comparisons by rank-based nonparametric Kruskal–Wallis test with FDR correction for multiple comparisons; *p<0.05, **p<0.01, ***p<0.001. Significant differences are possible when all medians are 0.000 because this is a rank-based test. Pairwise comparisons by Dunn's post-hoc tests for multiple comparisons (significant pairwise differences): [a]No breastfeeding (BF)/Partial BF with formula; [b]No BF/Partial BF without formula; [c]No BF/Exclusive BF; [d]Partial BF with formula/Partial BF without formula; [e]Partial BF with formula/Exclusive BF; no significant differences observed between Partial BF without formula and Exclusive BF. Breastfeeding refers to breast milk feeding, regardless of feeding mode (at the breast or from a bottle).

**eTable 8.** Median Relative Abundance of Abundant† Taxa in Fecal Microbiota of Infants at 12 Months According to Breastfeeding (BF) Duration[#]

| Phylum / *Family* | (A) Never Breastfed N=34 | (B) Breastfed < 6 months N=320 | (C) Breastfed 6 – 12 months N=149 | (D) Breastfed > 12 months N=316 | Overall pFDR | A vs B | A vs C | A vs D | B vs C | B vs D | C vs D |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Actinobacteria** | 1.010 | 2.060 | 1.667 | 3.637 | *** | | | a | | b | c |
| *Actinomycetaceae* | 0.008 | 0.008 | 0.008 | 0.008 | | | | | | | |
| *Bifidobacteriaceae* | 0.944 | 1.533 | 1.485 | 3.270 | *** | | | a | | b | c |
| *Coriobacteriaceae* | 0.129 | 0.117 | 0.086 | 0.086 | | | | | | | |
| *Micrococcaceae* | 0.000 | 0.000 | 0.000 | 0.000 | | | | | | | |
| **Bacteroidetes** | 50.845 | 48.509 | 52.810 | 47.002 | | | | | | | c |
| *Bacteroidaceae* | 46.682 | 40.262 | 47.045 | 42.293 | | | | | | | |
| *Porphyromonadaceae* | 0.183 | 0.121 | 0.055 | 0.008 | * | | | | | b | c |
| *Prevotellaceae* | 0.008 | 0.008 | 0.008 | 0.015 | | | | | | | |
| *Rikenellaceae* | 0.016 | 0.008 | 0.015 | 0.000 | *** | | | a | | b | c |
| **Firmicutes** | 33.699 | 37.739 | 33.623 | 36.673 | | | | | | | |
| Clostridiales (other) | 0.008 | 0.015 | 0.008 | 0.008 | *** | | | | | b | c |
| Clostridiales (unclassified) | 0.638 | 1.049 | 1.160 | 0.433 | *** | | | | | b | c |
| *Clostridiaceae* | 0.388 | 0.514 | 0.500 | 0.544 | | | | | | | |
| *Enterococcaceae* | 0.000 | 0.000 | 0.000 | 0.008 | ** | | | a | | b | c |
| *Erysipelotrichaceae* | 0.471 | 0.521 | 0.481 | 0.350 | ** | | | a | | b | c |
| *Gemellaceae* | 0.008 | 0.008 | 0.008 | 0.008 | | | | | | | |
| *Lachnospiraceae* | 14.006 | 14.595 | 14.457 | 12.861 | | | | | | b | |
| Lactobacillales (other) | 0.016 | 0.008 | 0.015 | 0.008 | | | | | | | |
| *Ruminococcaceae* | 6.624 | 7.715 | 6.682 | 4.272 | *** | | | | | b | c |
| *Streptococcaceae* | 0.277 | 0.283 | 0.281 | 0.319 | | | | | | | |
| *Veillonellaceae* | 1.925 | 3.944 | 3.685 | 8.749 | *** | | | a | | b | c |
| **Proteobacteria** | 3.042 | 3.869 | 3.962 | 5.694 | *** | | | a | | b | c |
| *Alcaligenaceae* | 0.012 | 0.704 | 0.721 | 0.956 | | | | a | | | |
| *Enterobacteriaceae* | 1.191 | 0.973 | 0.781 | 1.683 | * | | | | | b | c |
| *Pasteurellaceae* | 0.109 | 0.187 | 0.162 | 0.250 | | | | | | | |
| **Verrucomicrobia** | 0.016 | 0.031 | 0.016 | 0.008 | | | | | | b | |
| *Verrucomicrobiaceae* | 0.016 | 0.031 | 0.016 | 0.008 | | | | | | b | |
| Firmicutes/ Bacteroidetes ratio | 0.66 | 0.78 | 0.64 | 0.78 | | | | a | | b | c |

FDR, false discovery rate. †Taxa with median abundance >0% in 3 month and/or 12 month microbiota. [#]Feeding status at the time of sample collection. Overall comparisons by rank-based nonparametric Kruskal–Wallis test with FDR correction for multiple comparisons; *p<0.05, **p<0.01, ***p<0.001. Significant differences are possible when all medians are 0.000 because this is a rank-based test. Pairwise comparisons by Dunn's post-hoc tests for multiple comparisons (significant pairwise differences): [a]BF < 6 months/BF > 12 months; [b]BF 6 – 12 months/BF > 12 months; [c]Never BF/BF > 12 months; no significant differences were observed between BF <6 months/BF 6 – 12 months, Never BF/BF < 6 months or Never BF/BF 6 – 12 months. Breastfeeding refers to breast milk feeding, regardless of feeding mode (at the breast or from a bottle).

**eTable 9.** Median Relative Abundance of Abundant† Taxa in Fecal Microbiota of Infants at 3-4 and 12 Months According to Infant Weight Status at 12 Months

| Phylum / *Family* | Microbiota at 3-4 months (A) Normal | (B) At risk | (C) Overweight | Overall pFDR | Pairwise pFDR A vs B | A vs C | B vs C | Microbiota at 12 months (A) Normal | (B) At risk | (C) Overweight | Overall pFDR | Pairwise pFDR A vs B | A vs C | B vs C |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | N=699 | N=171 | N=67 | | | | | N=616 | N=137 | N=56 | | | | |
| **Actinobacteria** | 5.360 | 5.232 | 6.194 | | | | | 2.756 | 2.178 | 1.890 | | a | | c |
| *Actinomycetaceae* | 0.023 | 0.023 | 0.039 | | | | | 0.008 | 0.008 | 0.008 | | | | |
| *Bifidobacteriaceae* | 4.591 | 4.967 | 5.754 | | | | | 2.290 | 1.915 | 1.197 | | | | |
| *Coriobacteriaceae* | 0.039 | 0.055 | 0.085 | | | b | | 0.093 | 0.147 | 0.117 | | | | |
| *Micrococcaceae* | 0.008 | 0.008 | 0.015 | | | | | 0.000 | 0.000 | 0.000 | | | | |
| **Bacteroidetes** | 21.753 | 8.091 | 3.483 | | | | | 50.111 | 45.054 | 48.316 | | | | |
| *Bacteroidaceae* | 14.201 | 5.117 | 2.247 | | | | | 44.024 | 35.248 | 42.399 | | | | |
| *Porphyromonadaceae* | 0.008 | 0.008 | 0.008 | | | | | 0.023 | 0.078 | 0.012 | | | | |
| *Prevotellaceae* | 0.000 | 0.000 | 0.000 | | | | | 0.008 | 0.016 | 0.008 | | | | |
| *Rikenellaceae* | 0.000 | 0.000 | 0.000 | | | | | 0.008 | 0.008 | 0.008 | | | | |
| **Firmicutes** | 20.708 | 28.229 | 26.142 | | | | | 35.749 | 40.016 | 37.312 | | | | |
| Clostridiales (other) | 0.000 | 0.000 | 0.000 | | | | | 0.008 | 0.016 | 0.008 | | | | |
| Clostridiales (unclassified) | 0.000 | 0.000 | 0.000 | * | a | b | | 0.641 | 0.814 | 1.199 | | | | |
| *Clostridiaceae* | 0.330 | 0.574 | 0.581 | | | | | 0.516 | 0.564 | 0.385 | | | | |
| *Enterococcaceae* | 0.023 | 0.023 | 0.024 | | | | | 0.000 | 0.000 | 0.008 | | | | |
| *Erysipelotrichaceae* | 0.016 | 0.031 | 0.124 | | | b | | 0.425 | 0.499 | 0.356 | | | | |
| *Gemellaceae* | 0.000 | 0.000 | 0.008 | | | | | 0.008 | 0.008 | 0.008 | | a | | |
| *Lachnospiraceae* | 1.915 | 4.699 | 5.848 | * | a | b | | 13.392 | 15.163 | 14.450 | | | | |
| Lactobacillales (other) | 0.000 | 0.000 | 0.000 | | | | | 0.008 | 0.015 | 0.012 | | | | |
| *Ruminococcaceae* | 0.054 | 0.278 | 0.280 | | | b | | 6.362 | 6.460 | 6.454 | | | | |
| *Streptococcaceae* | 0.535 | 0.710 | 0.904 | | | | | 0.281 | 0.404 | 0.245 | | a | | c |
| *Veillonellaceae* | 4.226 | 6.266 | 5.537 | | | | | 5.250 | 5.814 | 3.978 | | | | |
| **Proteobacteria** | 18.196 | 19.921 | 22.815 | | | | | 4.448 | 3.887 | 5.758 | | | | |
| *Alcaligenaceae* | 0.000 | 0.000 | 0.000 | | | | | 0.748 | 0.598 | 0.733 | | | | |
| *Enterobacteriaceae* | 16.273 | 18.287 | 20.413 | | | | | 1.226 | 0.972 | 1.251 | | | | |
| *Pasteurellaceae* | 0.031 | 0.031 | 0.031 | | | | | 0.195 | 0.256 | 0.082 | | | | |
| **Verrucomicrobia** | 0.000 | 0.000 | 0.000 | | | | | 0.016 | 0.016 | 0.093 | | | | |
| *Verrucomicrobiaceae* | 0.000 | 0.000 | 0.000 | | | | | 0.016 | 0.016 | 0.093 | | | | |
| Firmicutes/ Bacteroidetes ratio | 1.29 | 3.08 | 11.36 | | | | | 0.73 | 0.84 | 0.76 | | | | |

FDR, false discovery rate. †Taxa with median abundance >0% in 3 month and/or 12 month microbiota. Overall comparisons by rank-based nonparametric Kruskal–Wallis test with FDR correction for multiple comparisons; *p<0.05, **p<0.01, ***p<0.001. Significant differences are possible when all medians are 0.000 because this is a rank-based test. Pairwise comparisons by Dunn's post-hoc tests for multiple comparisons (significant pairwise differences): [a]Normal/At Risk; [b]Normal/Overweight; [c]At Risk/Overweight

**eAppendix.** Detailed Methods

**DNA isolation**

Following collection and aliquotting, stool samples were maintained frozen at or below –80 C prior to analysis. For the isolation of community DNA, a modification of the Qiagen Isolation of DNA from Stool for Pathogen Detection protocol was used (Qiagen Inc., Valencia CA). In this procedure, a target mass of 200 mg (acceptable range from 80–220 mg, actual mass recorded) of frozen stool was combined with 1.4 mL Qiagen Stool Lysis Buffer (ASL), vortex mixed for 1 min or until the sample appeared thoroughly thawed and homogenized, and placed in a 95 C water bath for 5 min. Samples were then vortex mixed for 15 sec and centrifuged for 2 min at 14 Krpm. A volume of 1.2 mL of supernatant was removed to a new microcentrifuge tube, combined with a tablet of InhibitEX (Qiagen) and vortex mixed continuously for 1 min or until suspended. Samples were then incubated at room temperature for 1 min and centrifuged for 5 min at 14 Krpm. The remainder of the isolation procedure was carried out using a QIAcube robot following the "Pathogen Detection" program modified for a 60 µL elution volume rathewr than the standard 200 µL elution volume. Following completion of the program, DNA quality was evaluated by electrophoresis of 5 µL of isolated DNA in 1.0% agarose in 1×TAE buffer for 80 min at 80 VDV, visualized using SYBR safe stain (Thermo Fisher Scientific / Life Technologies Corp., Carlsbad CA) and recorded using a GelDoc XR+ Imaging system (BioRad Laboratories Inc., Hercules CA). DNA concentration was evaluated using a Quantifluor dsDNA system following the manufacturer's instructions, and adjusted to a final concentration of 5 ng/µL by the addition of 1×TE buffer.

**Amplification**

Bacterial 16S DNA from hypervariable region V4 was amplified by PCR using the core forward primer V4+515F (5'-TATGGTAATTGTGTGCCAGCMGCCGCGGTAA-3') and the core reverse primer V4-806R (5'-AGTCAGTCAGCCGGACTACHVGGGTWTCTAAT-3') synthesized to include Illumina adapers, primer pad and linker sequences, a Golay barcode (forward primer). PCRs were conducted in a final volume of 25 µL, consisting of 10 ng template DNA and 0.6 µM of each primer in Kapa2G Robust Hotstart Taq ready mix (KapaBiosystems, Wilmington MA) at 1× concentration. PCR conditions consisted of an initial denaturation at 94 for 3 min followed by 20 cycles of 94 C for 30 sec (denaturation), 50 C for 30 sec (annealing), and 72 C for 30 sec (extension), with a completion step at 10 C. Amplicon quality was assessed by electrophoresis following the procedure described in the previous section and quantitated based on the summed brightness of bands as determined by the GelDoc XR+ analyser. Samples yielding less 100 ng of total product were reamplified using diluted template DNA to reduce the concentration of PCR inhibitors. PCR products were combined for multiplex sequencing in batches of 48 up to a maximum of 96. Total volume was reduced using Amicon Ultra centrifugal filter concentrators (Millipore Sigma, Burlington MA) to between 25–50 µL. Concentrated products electrophoresed on a 1.4% agarose gel in 1×TAE containing SYBR safe stain at 80 VDC for 90 min. Bands were excised and purified using a GeneClean Turbo Kit (MP Biomedicals, Santa Ana, CA) following the manufacturer's directions modified to elute the product in 40 µL DES for 10 min followed by centrifugation of final eluent for 2 min. Final DNA concentration was determined by Quantifluor, as above.

**Sequencing and bioinformatics**

Samples were sequenced by Illumina MiSeq (San Diego, CA) using 150 bp paired-end (x2) V2 chemistry. Data were outputted in a format consisting of two fastq files containing read 1 and read 2 datasets, and a third fastq file corresponding to barcodes. Following decompression and concatenation of data files, forward and reverse assembled and the resultant contigs binned by barcode using the Qiime (version 1.9.1) open-source bioinformatics pipeline. The analysis pipeline, in brief, consisted of the following; Non-bacterial sequences were excluded as those that failed to cluster against the Greengenes reference database (version May 2013) at 60% similarity. The resulting filtered dataset was subjected to closed reference picking against the Greengenes reference database at 97% similarity using USEARCH6.1. Sequences that failed to cluster were aggregated over the entire dataset (singletons removed) and subjected to de novo clustering using USEARCH10 (64 bit). Taxonomies were assigned according to the Greengenes reference database, and closed- and de novo picked datasets were merged.

**Pipeline command sequence summary**

**#MiSeq FASTQ FROM SEQUENCER - UNZIPPING AND CONCATENATING FILES**
**#note used all files instead of just the 'undefined' files and allowed the quality parameters to cull for consistency across runs**
gunzip *I1*.gz
gunzip *R1*.gz
gunzip *R2*.gz
cat *S1_L001_I1_001.fastq *S2_L001_I1_001.fastq *S3_L001_I1_001.fastq *S4_L001_I1_001.fastq *S5_L001_I1_001.fastq
*S6_L001_I1_001.fastq *S7_L001_I1_001.fastq *S8_L001_I1_001.fastq *S9_L001_I1_001.fastq *S10_L001_I1_001.fastq
*S11_L001_I1_001.fastq *S12_L001_I1_001.fastq Undetermined_S0_L001_I1_001.fastq > cat_index.fastq
cat *S1_L001_R1_001.fastq *S2_L001_R1_001.fastq *S3_L001_R1_001.fastq *S4_L001_R1_001.fastq *S5_L001_R1_001.fastq
*S6_L001_R1_001.fastq *S7_L001_R1_001.fastq *S8_L001_R1_001.fastq *S9_L001_R1_001.fastq *S10_L001_R1_001.fastq
*S11_L001_R1_001.fastq *S12_L001_R1_001.fastq Undetermined_S0_L001_R1_001.fastq > cat_R1.fastq

```
cat *S1_L001_R2_001.fastq *S2_L001_R2_001.fastq *S3_L001_R2_001.fastq *S4_L001_R2_001.fastq *S5_L001_R2_001.fastq
*S6_L001_R2_001.fastq *S7_L001_R2_001.fastq *S8_L001_R2_001.fastq *S9_L001_R2_001.fastq *S10_L001_R2_001.fastq
*S11_L001_R2_001.fastq *S12_L001_R2_001.fastq Undetermined_S0_L001_R2_001.fastq > cat_R2.fastq
```
#_____

**#MiSeq RUN CLOSED REFERENCE OTU-PICKING PIPELINE**
**#reference database = Greengenes version May 2013**
**#count the number of reads for input from MiSeq R1 fastq file**
count_seqs.py -i cat_R1.fastq -o cat_R1_seq_count.txt
**#assembling forward and reverse reads**
join_paired_ends.py -m fastq-join -j 25 -p 5 -b cat_index.fastq -f cat_R1.fastq -r cat_R2.fastq -o fastq_join/
**#binning sequences by bar code**
split_libraries_fastq.py -i fastq_join/fastqjoin.join.fastq -b fastq_join/fastqjoin.join_barcodes.fastq -m mapping_file.txt -r 3 -p 0.00 -n 0 --
rev_comp_mapping_barcodes --barcode_type 12 --max_barcode_errors 1.5 -o split_seqs/
**#identify non-bacterial reads to be filtered**
parallel_pick_otus_usearch61_ref.py -i split_seqs/seqs.fna -r /home/james/qiime_software/gg_13_8_otus/rep_set/97_otus.fasta --
usearch61_sort_method abundance --sizeorder --similarity 0.6 --jobs_to_start 16 -o prefilter/
**#count the number of sequences that failed to cluster with reference data set at 60%**
wc -l prefilter/seqs_failures.txt > prefilter/seqs_failures_count.txt
**#create filtered data set**
filter_fasta.py -f split_seqs/seqs.fna -s prefilter/seqs_failures.txt -n -o prefilter/prefiltered_seqs.fna
**#parallel closed reference pick OTUs using usearch61**
parallel_pick_otus_usearch61_ref.py -i prefilter/prefiltered_seqs.fna -r /home/james/qiime_software/gg_13_8_otus/rep_set/97_otus.fasta -
-usearch61_sort_method abundance --sizeorder --similarity 0.97 --jobs_to_start 16 -o closed_ref_OTUs/
**#count the number of sequences that failed to cluster with reference data set at 97%**
wc -l closed_ref_OTUs/prefiltered_seqs_failures.txt > closed_ref_OTUs/prefiltered_seqs_failures_count.txt
**#pick representative sequence from each OTU cluster**
pick_rep_set.py -i closed_ref_OTUs/prefiltered_seqs_otus.txt -f prefilter/prefiltered_seqs.fna -o closed_ref_OTUs/rep_set.fasta
**#Assign taxonomy using uclust**
assign_taxonomy.py -i closed_ref_OTUs/rep_set.fasta -m uclust -r /home/james/qiime_software/gg_13_8_otus/rep_set/97_otus.fasta -t
/home/james/qiime_software/gg_13_8_otus/taxonomy/97_otu_taxonomy.txt --uclust_min_consensus_fraction 0.51 --uclust_similarity 0.9
--uclust_max_accepts 3 -o closed_ref_OTUs/uclust_closed_ref_tax_assign/
**#make biom formatted OTU table**
make_otu_table.py -i closed_ref_OTUs/prefiltered_seqs_otus.txt -t
closed_ref_OTUs/uclust_closed_ref_tax_assign/rep_set_tax_assignments.txt -o
closed_ref_OTUs/uclust_closed_ref_tax_assign/uclust_closed_ref_picked_OTUs.biom
**#summarize biom table**
biom summarize-table -i closed_ref_OTUs/uclust_closed_ref_tax_assign/uclust_closed_ref_picked_OTUs.biom -o
closed_ref_OTUs/uclust_closed_ref_tax_assign/uclust_closed_ref_picked_OTUs_biom_table_summary.txt
**#count the number of OTUs per sample**
alpha_diversity.py -i closed_ref_OTUs/uclust_closed_ref_tax_assign/uclust_closed_ref_picked_OTUs.biom -m observed_species -o
closed_ref_OTUs/uclust_closed_ref_tax_assign/uclust_closed_ref_picked_OTU_count_per_sample.txt
**#convert biom-formatted table to tab-delimited text format**
**#note that this table is not used in downstream pipeline**
biom convert -i closed_ref_OTUs/uclust_closed_ref_tax_assign/uclust_closed_ref_picked_OTUs.biom -o
closed_ref_OTUs/uclust_closed_ref_tax_assign/uclust_closed_ref_picked_OTUs.txt -b --header-key taxonomy
**#make directory for de novo picking files**
mkdir denovo_OTUs
**#pull reads that failed to cluster with reference database for de novo OTU picking**
filter_fasta.py -f prefilter/prefiltered_seqs.fna -s closed_ref_OTUs/prefiltered_seqs_failures.txt -o
denovo_OTUs/seqs_for_denovo_pick.fna
**#count the number of reads for input into denovo-picking step (note will/should be the same number as output in**
**prefiltered_seqs_failures_count.txt)**
count_seqs.py -i denovo_OTUs/seqs_for_denovo_pick.fna -o denovo_OTUs/seq_count.txt
#_____

**# merge biom tables from sym closed OTU picking and characterize 1) full table 2) table with singletons removed**
merge_otu_tables.py -i
sym1_130211_uclust_closed_ref_picked_OTUs.biom,sym2_130212_uclust_closed_ref_picked_OTUs.biom,sym3_130214_uclust_closed
_ref_picked_OTUs.biom,sym4_130221_uclust_closed_ref_picked_OTUs.biom,sym5_130226_uclust_closed_ref_picked_OTUs.biom,sy
m6_130227_uclust_closed_ref_picked_OTUs.biom,sym7_130304_uclust_closed_ref_picked_OTUs.biom,sym8_130306_uclust_closed_r
ef_picked_OTUs.biom,sym9_130416_uclust_closed_ref_picked_OTUs.biom,sym10_130418_uclust_closed_ref_picked_OTUs.biom,sym
11_130423_uclust_closed_ref_picked_OTUs.biom,sym12_130424_uclust_closed_ref_picked_OTUs.biom,sym13_130429_uclust_closed
_ref_picked_OTUs.biom,sym14_130502_uclust_closed_ref_picked_OTUs.biom,sym15_130506_uclust_closed_ref_picked_OTUs.biom,s

ym16_130417_uclust_closed_ref_picked_OTUs.biom,sym20_130711_uclust_closed_ref_picked_OTUs.biom,sym21_130716_uclust_clo
sed_ref_picked_OTUs.biom,sym22_130717_uclust_closed_ref_picked_OTUs.biom,sym23_130718_uclust_closed_ref_picked_OTUs.bio
m,sym24_130924_uclust_closed_ref_picked_OTUs.biom,sym25_130925_uclust_closed_ref_picked_OTUs.biom,sym26_130930_uclust_
closed_ref_picked_OTUs.biom,sym27_131001_uclust_closed_ref_picked_OTUs.biom,sym28_131009_uclust_closed_ref_picked_OTUs
_corrected.biom,sym29_131106_uclust_closed_ref_picked_OTUs.biom,sym30_131111_uclust_closed_ref_picked_OTUs_corrected.biom
,sym31_140225_uclust_closed_ref_picked_OTUs.biom,sym33_140219_uclust_closed_ref_picked_OTUs.biom,sym34_140220_uclust_cl
osed_ref_picked_OTUs.biom,sym35_140227_uclust_closed_ref_picked_OTUs.biom,sym36_140916_uclust_closed_ref_picked_OTUs.bi
om,sym37_140918_uclust_closed_ref_picked_OTUs.biom,sym38_140818_uclust_closed_ref_picked_OTUs.biom,sym39_140826_uclust
_closed_ref_picked_OTUs.biom,sym40_140826_uclust_closed_ref_picked_OTUs_corrected.biom,sym41_140818_uclust_closed_ref_pic
ked_OTUs.biom,sym42_141103_uclust_closed_ref_picked_OTUs.biom,sym43_141103_uclust_closed_ref_picked_OTUs.biom,sym44_1
40916_uclust_closed_ref_picked_OTUs.biom,sym45_150113_uclust_closed_ref_picked_OTUs.biom,sym46_150114_uclust_closed_ref_
picked_OTUs.biom,sym47_150224_uclust_closed_ref_picked_OTUs.biom,sym48_150225_uclust_closed_ref_picked_OTUs.biom,sym4
9_150225_uclust_closed_ref_picked_OTUs.biom,sym50_150504_uclust_closed_ref_picked_OTUs.biom,sym51_150611_uclust_closed_
ref_picked_OTUs.biom,sym52_150506_uclust_closed_ref_picked_OTUs.biom,sym53_150507_uclust_closed_ref_picked_OTUs.biom,sy
m54_150602_uclust_closed_ref_picked_OTUs.biom,sym55_150603_uclust_closed_ref_picked_OTUs.biom,sym56_150604_uclust_close
d_ref_picked_OTUs.biom,sym57_150604_uclust_closed_ref_picked_OTUs.biom,sym58_150707_uclust_closed_ref_picked_OTUs.biom
,sym59_150708_uclust_closed_ref_picked_OTUs.biom,sym60_150709_uclust_closed_ref_picked_OTUs.biom,sym61_150715_uclust_cl
osed_ref_picked_OTUs.biom,sym62_150817_uclust_closed_ref_picked_OTUs.biom,sym63_150812_uclust_closed_ref_picked_OTUs.bi
om,sym64_150818_uclust_closed_ref_picked_OTUs.biom,sym65_150819_uclust_closed_ref_picked_OTUs.biom,sym66_151001_uclust
_closed_ref_picked_OTUs.biom,sym67_151006_uclust_closed_ref_picked_OTUs.biom,sym68_160105_uclust_closed_ref_picked_OTU
s.biom,sym69_151007_uclust_closed_ref_picked_OTUs.biom,sym70_151109_uclust_closed_ref_picked_OTUs.biom,sym71_151110_uc
lust_closed_ref_picked_OTUs.biom,sym72_151111_uclust_closed_ref_picked_OTUs.biom,sym73_151125_uclust_closed_ref_picked_O
TUs.biom,sym74_151209_uclust_closed_ref_picked_OTUs.biom,sym75_151210_uclust_closed_ref_picked_OTUs.biom,sym76_151214
_uclust_closed_ref_picked_OTUs.biom,sym77_151215_uclust_closed_ref_picked_OTUs.biom,sym78_161114_uclust_closed_ref_picked
_OTUs.biom,sym79_160216_uclust_closed_ref_picked_OTUs.biom,sym80_160217_uclust_closed_ref_picked_OTUs.biom,sym81_1602
17_uclust_closed_ref_picked_OTUs.biom,sym82_160329_uclust_closed_ref_picked_OTUs.biom,sym83_160329_uclust_closed_ref_pic
ked_OTUs.biom,sym84_160404_uclust_closed_ref_picked_OTUs.biom,sym85_160404_uclust_closed_ref_picked_OTUs.biom,sym86_1
60629_uclust_closed_ref_picked_OTUs.biom,sym87_160718_uclust_closed_ref_picked_OTUs.biom,sym88_160705_uclust_closed_ref_
picked_OTUs.biom,sym89_160707_uclust_closed_ref_picked_OTUs.biom,sym90_160809_uclust_closed_ref_picked_OTUs.biom,sym9
1_160809_uclust_closed_ref_picked_OTUs.biom,sym92_160825_uclust_closed_ref_picked_OTUs.biom,sym93_160825_uclust_closed_
ref_picked_OTUs.biom,sym94_160928_uclust_closed_ref_picked_OTUs.biom,sym95_160929_uclust_closed_ref_picked_OTUs.biom,sy
m96_160929_uclust_closed_ref_picked_OTUs.biom,sym97_160929_uclust_closed_ref_picked_OTUs.biom,sym98_161012_uclust_close
d_ref_picked_OTUs.biom,sym99_161012_uclust_closed_ref_picked_OTUs.biom,sym100_161013_uclust_closed_ref_picked_OTUs.bio
m -o merged_symbiota_closed_picked_table_1to100.biom

**# characterize full biom table**
biom summarize-table -i merged_symbiota_closed_picked_table_1to100.biom -o
merged_symbiota_closed_picked_table_1to100_biom_table_summary.txt
alpha_diversity.py -i merged_symbiota_closed_picked_table_1to100.biom -m observed_species -o
merged_symbiota_closed_picked_table_1to100_OTU_count_per_sample.txt
#_____

**#COMBINED MiSeq RUN DENOVO OTU-PICKING PIPELINE USING USEARCH10 64 bit**
**# concatenate reads that failed to cluster with reference database into single fasta file for denovo picking**
cat sym1_130211_seqs_for_denovo_pick.fna sym2_130212_seqs_for_denovo_pick.fna sym3_130214_seqs_for_denovo_pick.fna
sym4_130221_seqs_for_denovo_pick.fna sym5_130226_seqs_for_denovo_pick.fna sym6_130227_seqs_for_denovo_pick.fna
sym7_130304_seqs_for_denovo_pick.fna sym8_130306_seqs_for_denovo_pick.fna sym9_130416_seqs_for_denovo_pick.fna
sym10_130418_seqs_for_denovo_pick.fna sym11_130423_seqs_for_denovo_pick.fna sym12_130424_seqs_for_denovo_pick.fna
sym13_130429_seqs_for_denovo_pick.fna sym14_130502_seqs_for_denovo_pick.fna sym15_130506_seqs_for_denovo_pick.fna
sym16_130417_seqs_for_denovo_pick.fna sym20_130711_seqs_for_denovo_pick.fna sym21_130716_seqs_for_denovo_pick.fna
sym22_130717_seqs_for_denovo_pick.fna sym23_130718_seqs_for_denovo_pick.fna sym24_130924_seqs_for_denovo_pick.fna
sym25_130925_seqs_for_denovo_pick.fna sym26_130930_seqs_for_denovo_pick.fna sym27_131001_seqs_for_denovo_pick.fna
sym28_131009_corrected_seqs_for_denovo_pick.fna sym29_131106_seqs_for_denovo_pick.fna
sym30_131111_corrected_seqs_for_denovo_pick.fna sym31_140225_seqs_for_denovo_pick.fna
sym33_140219_seqs_for_denovo_pick.fna sym34_140220_seqs_for_denovo_pick.fna sym35_140227_seqs_for_denovo_pick.fna
sym36_140916_seqs_for_denovo_pick.fna sym37_140918_seqs_for_denovo_pick.fna sym38_140818_seqs_for_denovo_pick.fna
sym39_140826_seqs_for_denovo_pick.fna sym40_140826_corrected_seqs_for_denovo_pick.fna
sym41_140818_seqs_for_denovo_pick.fna sym42_141103_seqs_for_denovo_pick.fna sym43_141103_seqs_for_denovo_pick.fna
sym44_140916_seqs_for_denovo_pick.fna sym45_150113_seqs_for_denovo_pick.fna sym46_150114_seqs_for_denovo_pick.fna
sym47_150224_seqs_for_denovo_pick.fna sym48_150225_seqs_for_denovo_pick.fna sym49_150225_seqs_for_denovo_pick.fna
sym50_150504_seqs_for_denovo_pick.fna sym51_150611_seqs_for_denovo_pick.fna sym52_150506_seqs_for_denovo_pick.fna
sym53_150507_seqs_for_denovo_pick.fna sym54_150602_seqs_for_denovo_pick.fna sym55_150603_seqs_for_denovo_pick.fna
sym56_150604_seqs_for_denovo_pick.fna sym57_150604_seqs_for_denovo_pick.fna sym58_150707_seqs_for_denovo_pick.fna
sym59_150708_seqs_for_denovo_pick.fna sym60_150709_seqs_for_denovo_pick.fna sym61_150715_seqs_for_denovo_pick.fna

sym62_150817_seqs_for_denovo_pick.fna sym63_150812_seqs_for_denovo_pick.fna sym64_150818_seqs_for_denovo_pick.fna
sym65_150819_seqs_for_denovo_pick.fna sym66_151001_seqs_for_denovo_pick.fna sym67_151006_seqs_for_denovo_pick.fna
sym68_160105_seqs_for_denovo_pick.fna sym69_151007_seqs_for_denovo_pick.fna sym70_151109_seqs_for_denovo_pick.fna
sym71_151110_seqs_for_denovo_pick.fna sym72_151111_seqs_for_denovo_pick.fna sym73_151125_seqs_for_denovo_pick.fna
sym74_151209_seqs_for_denovo_pick.fna sym75_151210_seqs_for_denovo_pick.fna sym76_151214_seqs_for_denovo_pick.fna
sym77_151215_seqs_for_denovo_pick.fna sym78_xxxxxx_seqs_for_denovo_pick.fna sym79_151215_seqs_for_denovo_pick.fna
sym80_160217_seqs_for_denovo_pick.fna sym81_160217_seqs_for_denovo_pick.fna sym82_160329_seqs_for_denovo_pick.fna
sym83_160329_seqs_for_denovo_pick.fna sym84_160404_seqs_for_denovo_pick.fna sym85_160404_seqs_for_denovo_pick.fna
sym86_160629_seqs_for_denovo_pick.fna sym87_160718_seqs_for_denovo_pick.fna sym88_160705_seqs_for_denovo_pick.fna
sym89_160707_seqs_for_denovo_pick.fna sym90_160809_seqs_for_denovo_pick.fna sym91_160809_seqs_for_denovo_pick.fna
sym92_160825_seqs_for_denovo_pick.fna sym93_160825_seqs_for_denovo_pick.fna sym94_160928_seqs_for_denovo_pick.fna
sym95_160929_seqs_for_denovo_pick.fna sym96_160929_seqs_for_denovo_pick.fna sym97_160929_seqs_for_denovo_pick.fna
sym98_161012_seqs_for_denovo_pick.fna sym99_161012_seqs_for_denovo_pick.fna sym100_161013_seqs_for_denovo_pick.fna >
concat_sym_1to100_seqs_for_denovo_pick.fna
count_seqs.py -i concat_sym_1to100_seqs_for_denovo_pick.fna -o concat_sym_1to100_seqs_for_denovo_pick_fna_seq_count.txt
**# use filter command to relabel with sample number**
**# get unique sequences for clustering**
usearch10 -fastx_uniques cat_sym_1to100_reads_for_denovo_pick.fna -sizeout -fastaout
cat_sym_1to100_reads_for_denovo_pick_Uniqs.fa
        00:44 4.0Gb  100.0% Reading cat_sym_1to100_reads_for_denovo_pick.fna
        00:44 3.9Gb  CPU has 16 cores, defaulting to 10 threads
        00:50 6.2Gb  100.0% DF
        00:51 6.3Gb  9382875 seqs, 7658500 uniques, 7331721 singletons (95.7%)
        00:51 6.3Gb  Min size 1, median 1, max 105358, avg 1.23
        01:56 4.8Gb  100.0% Writing cat_sym_1to100_reads_for_denovo_pick_Uniqs.fa
**# cluster using usearch10 with n=1 - ONLY FOR TAXON ASSIGNMENT -- TOO LARGE TO MAKE TABLE**
usearch10 -cluster_otus cat_sym_1to100_reads_for_denovo_pick_Uniqs.fa -minsize 2 -relabel OTU -otus
cat_sym_1to100_reads_for_denovo_pick_Uniqs_OTUrepSet.fa
        05:13 89Mb   100.0% 5869 OTUs, 46132 chimeras
**# assign names with Qiime against Greengenes for consistency with closed picked OTUs (note: need to attach these to OTUs in table using biom command)**
assign_taxonomy.py -i cat_sym_1to100_reads_for_denovo_pick_Uniqs_OTUrepSet.fa -m uclust -r
/home/james/qiime_software/gg_13_8_otus/rep_set/97_otus.fasta -t
/home/james/qiime_software/gg_13_8_otus/taxonomy/97_otu_taxonomy.txt --uclust_min_consensus_fraction 0.51 --uclust_similarity 0.9
--uclust_max_accepts 3 -o uclust_tax_assign/
**# cluster using usearch10 with n=100**
usearch10 -cluster_otus cat_sym_1to100_reads_for_denovo_pick_Uniqs.fa -minsize 100 -relabel OTU -otus
cat_sym_1to100_reads_for_denovo_pick_Uniqs_OTUrepSet_n10.fa
usearch10 -otutab cat_sym_1to100_reads_for_denovo_pick.fna -otus
cat_sym_1to100_reads_for_denovo_pick_Uniqs_OTUrepSet_n100.fa -biomout cat_sym_1to100_denovoOTUtable.json -mapout
denovoOTUs_map.txt -notmatched denovo_unmapped_reads.fa -dbmatched denovo_matched_reads_with_sizes.fa -sizeout
usearch10 -otutab cat_sym_1to100_reads_for_denovo_pick.fna -otus
cat_sym_1to100_reads_for_denovo_pick_Uniqs_OTUrepSet_n100.fa -otutabout cat_sym_1to100_denovoOTUtable_n100.txt
# _____
biom convert -i cat_sym_1to100_denovoOTUtable_n100.txt -o cat_sym_1to100_denovoOTUtable_n100.biom --table-type='OTU table' --
process-obs-metadata taxonomy
biom add-metadata --sc-separated taxonomy --observation-header OTUID,taxonomy --observation-metadata-fp
cat_sym_1to100_reads_for_denovo_pick_Uniqs_OTUrepSet_tax_assignments.txt -i cat_sym_1to100_denovoOTUtable_n100.biom -o
cat_sym_1to100_denovoOTUtable_n100_wTaxa.biom
biom summarize-table -i cat_sym_1to100_denovoOTUtable_n100_wTaxa.biom -o
cat_sym_1to100_denovoOTUtable_n100_wTaxa_biom_summary.txt
**# make .txt version of table to check labels etc**
biom convert -i cat_sym_1to100_denovoOTUtable_n100_wTaxa.biom -o cat_sym_1to100_denovoOTUtable_n100_wTaxa_biom.txt -b --
header-key taxonomy
# _____
**# filter closed pick table at n100 to match**
filter_otus_from_otu_table.py -i merged_symbiota_closed_picked_table_1to100.biom -n 100 -o
merged_symbiota_closed_picked_table_1to100_n100.biom
**# make text version of table to quality check**
biom convert -i merged_symbiota_closed_picked_table_1to100_n100.biom -o
merged_symbiota_closed_picked_table_1to100_n100_biom_.txt -b --header-key taxonomy

biom summarize-table -i merged_symbiota_closed_picked_table_1to100_n100.biom -o
merged_symbiota_closed_picked_table_1to100_n100_biom_summary.txt
# _____
**# merge closed picked and denovo picked tables**
merge_otu_tables.py -i
cat_sym_1to100_denovoOTUtable_n100_wTaxa.biom,merged_symbiota_closed_picked_table_1to100_n100.biom -o
FINAL_merged_closed_and_denovo_picked_OTUtable_filter100.biom
biom summarize-table -i FINAL_merged_closed_and_denovo_picked_OTUtable_filter100.biom -o
FINAL_merged_closed_and_denovo_picked_OTUtable_filter100_biom_summary.txt
biom convert -i merged_closed_denovo_sym1to100_Feb12_2017.biom -o merged_closed_denovo_sym1to100_Feb12_2017_biom.txt -b -
-header-key taxonomy