Supplementary figures

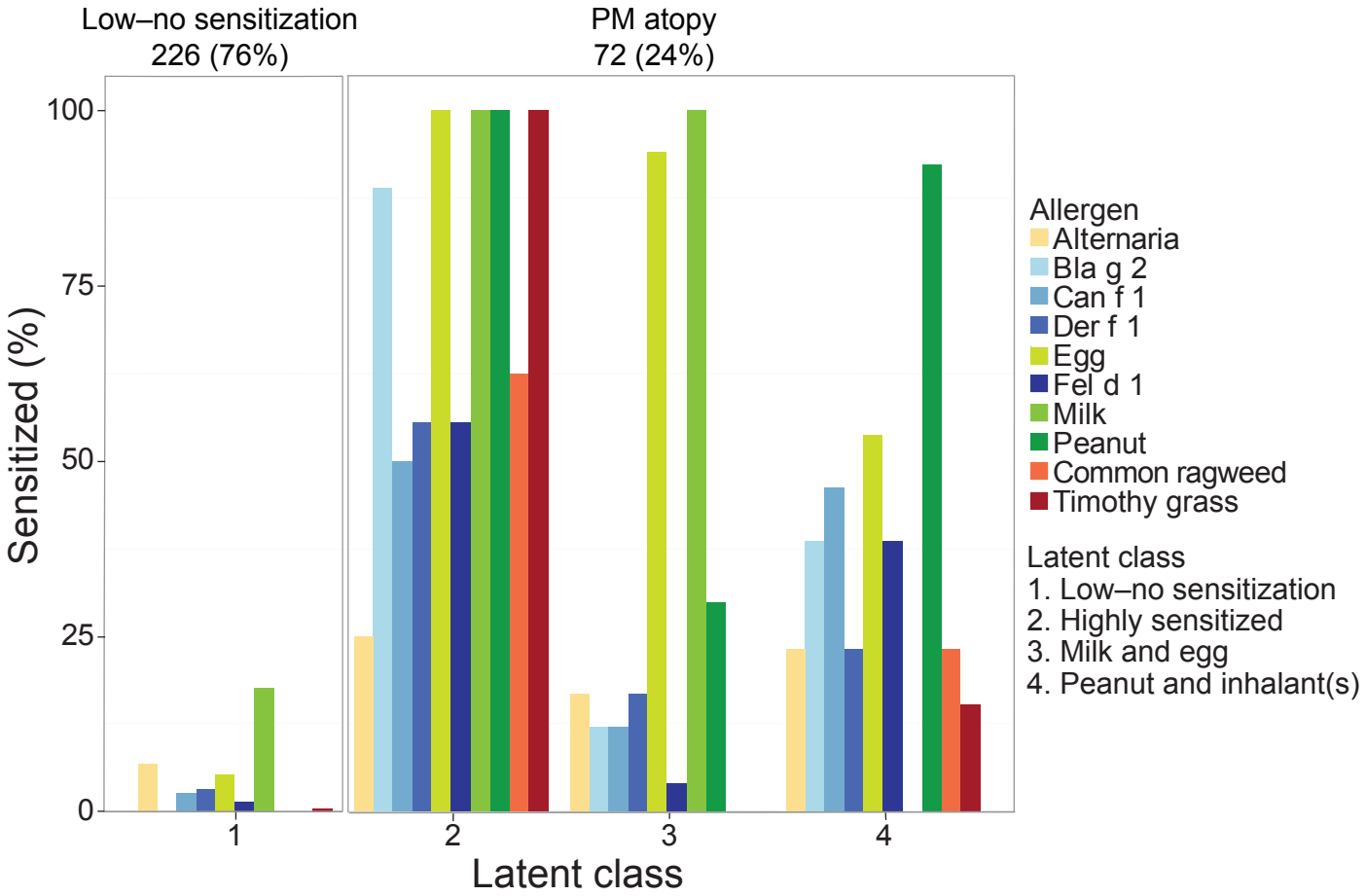

Fig. 1. Atopy status at age two–years is defined by latent class analysis, which clusters participants into one of four groups based on their pattern of specific IgE response to 10 common allergens. For the purpose of this study, latent classes 2,3 and 4 were grouped as the predominately multi–sensitized (PM atopy) group. Number of participants in each group and percent of population is provided.

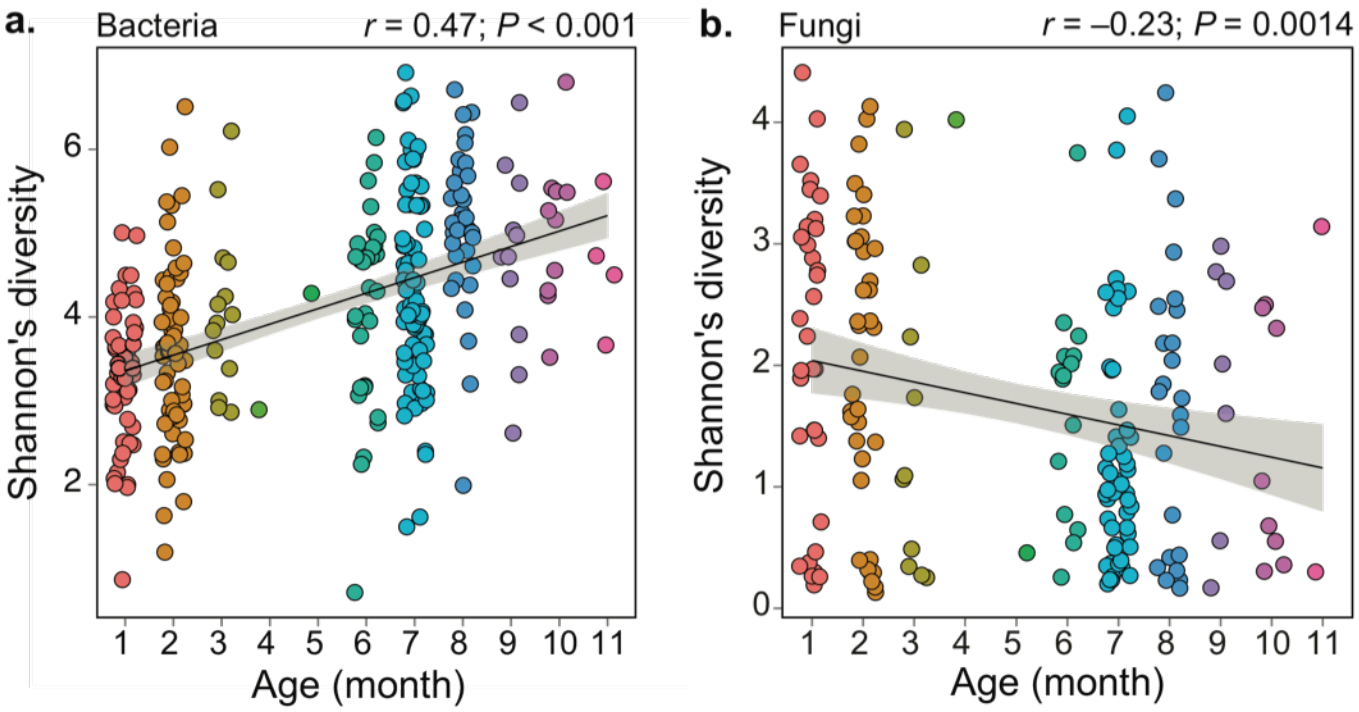

Fig. 2. Bacterial and fungal α–diversity correlate with age of participant at the time of stool sample collection. (**a**) Bacterial diversity positively correlated with increasing age at the time of stool sample collection (*n* = 298; Pearson's correlation; *r* = 0.47; *P* < 0.001). (**b**) Fungal diversity negatively correlated with increasing age of stool sample collection (*n* = 188; Pearson's correlation; *r* = –0.23; *P* = 0.0014).

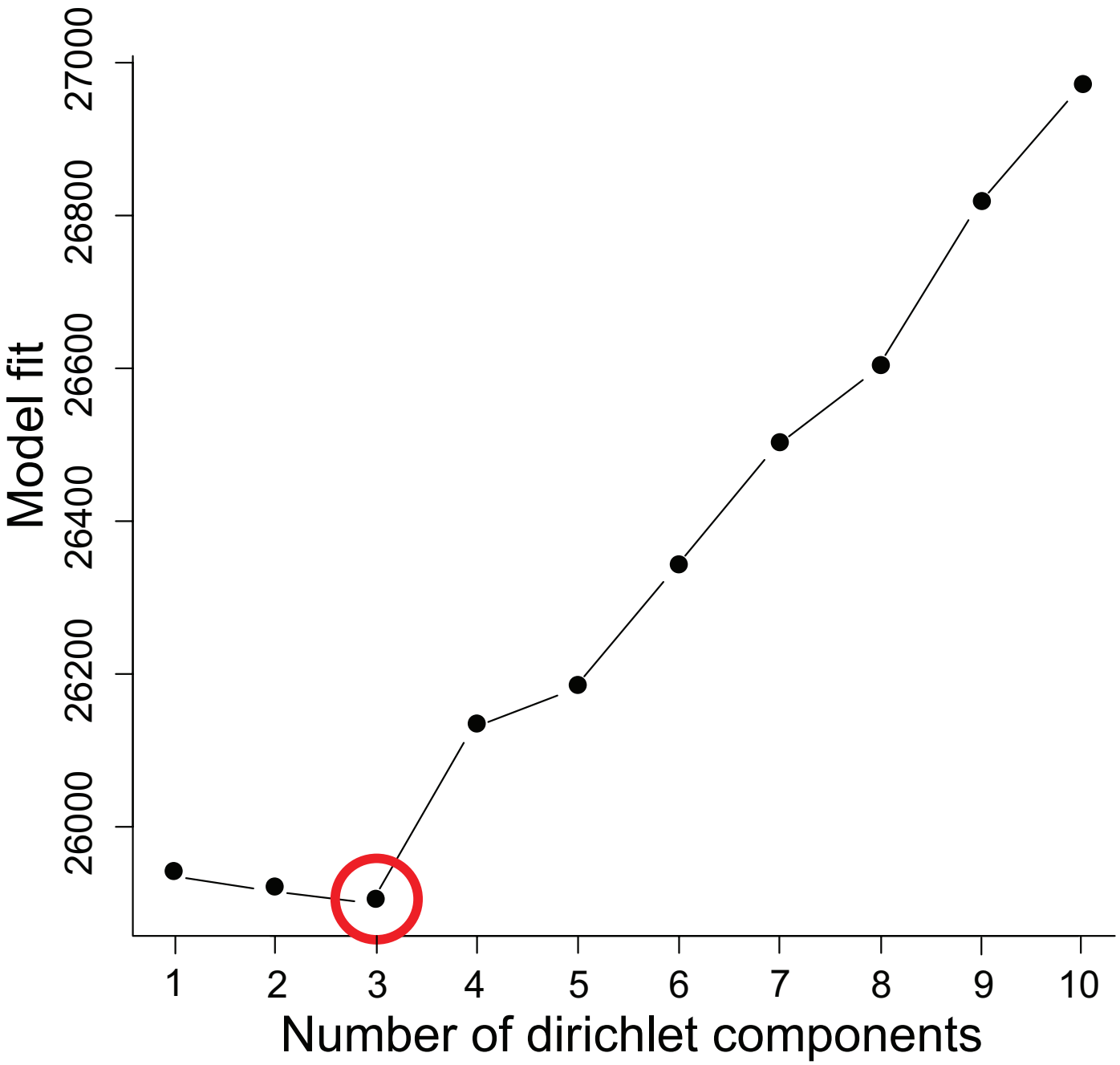

Fig. 3 Dirichlet multinomial mixture model identifies three compositionally distinct NGMs as the best model fit. Model fit was based on the Laplace approximation to the negative log model where a lower value indicates a better model fit.

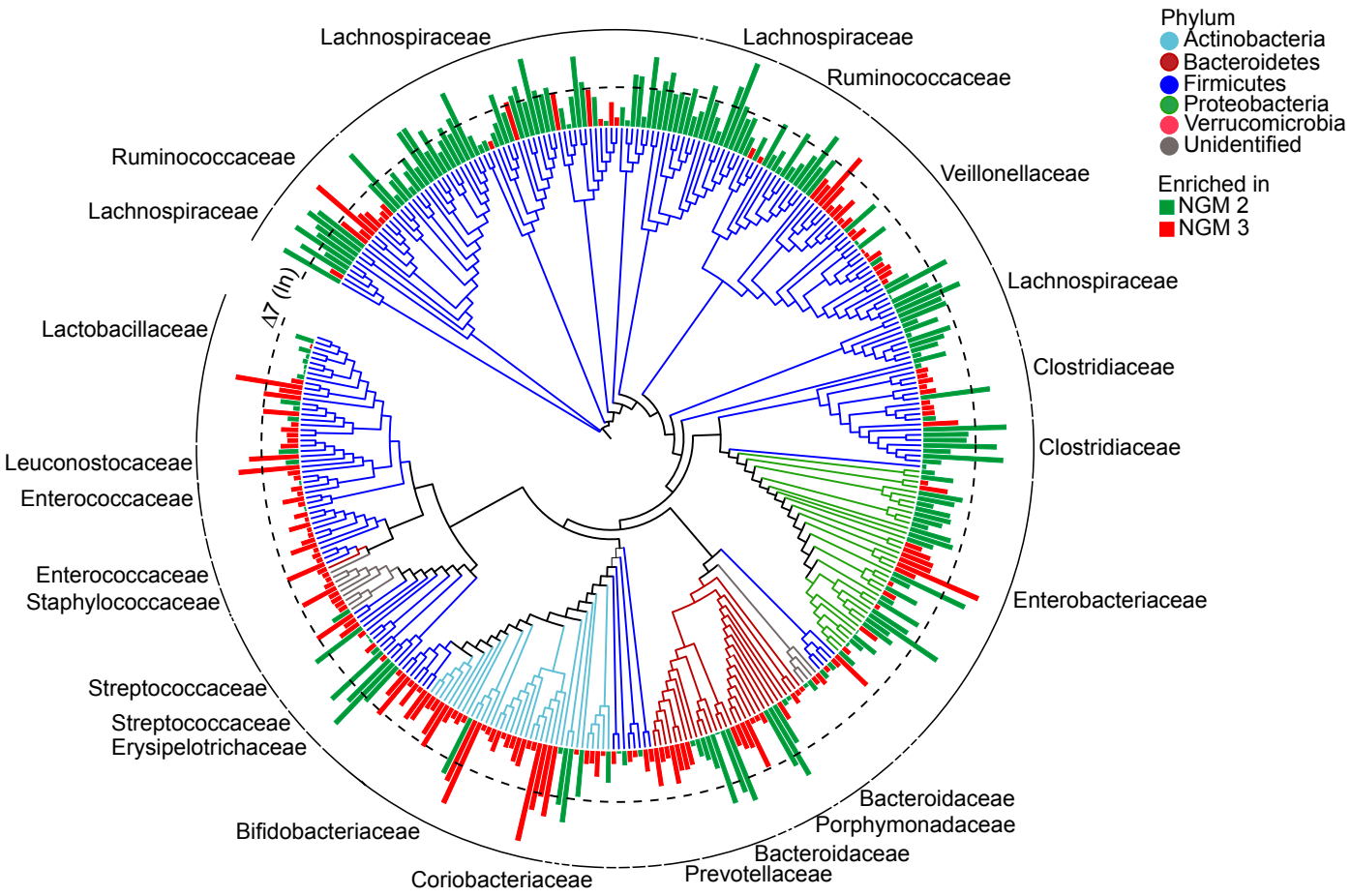

Fig. 4. **NGM2 and NGM3 gut microbiota exhibit significant differences in bacterial taxonomic content.** Zero–inflated negative binomial regression model corrected using Benjamini–Hochberg method for false discovery, identified taxa present in significantly different relative abundance between NGM2 and NGM3, $q < 0.05$ ($n = 130$). Relative abundance deltas were natural log–transformed prior to plotting on phylogenetic tree. Height of bars indicates the magnitude of relative abundance delta across comparator groups.

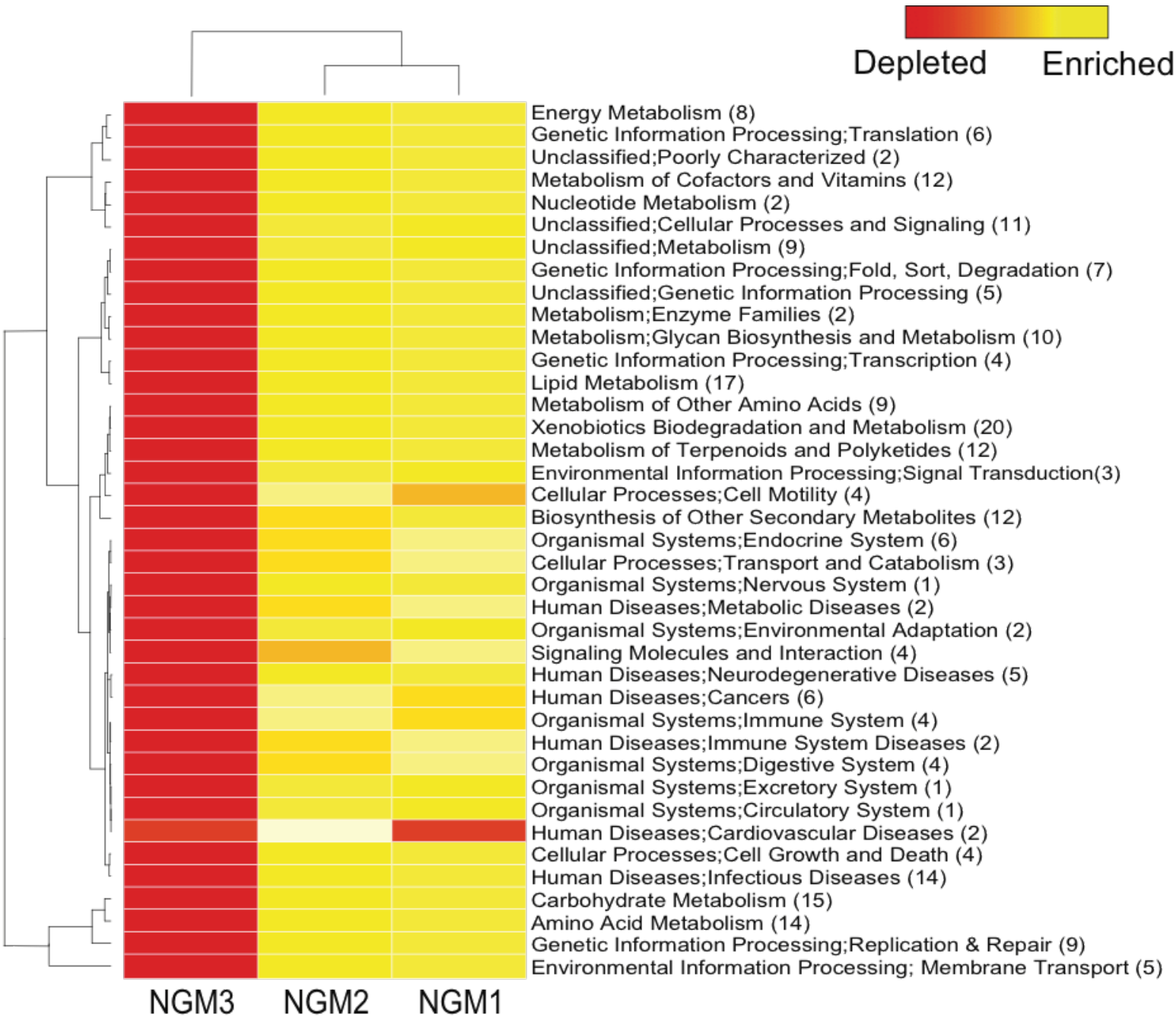

Fig. 5. Comparison of PICRUSt predicted bacterial pathways encoded by taxa significantly differentially enriched or depleted across NGM1 and 2, compared with NGM3. *In silico* metagenomic predictions were based specifically on those taxa that significantly differentiated the three microbiota–states based on zero–inflated negative binomial regression (n = 130). Bacterial amino acid (n = 23), xenobiotic (n = 20) and lipid (n = 17) metabolism pathways represented a large proportion of bacterial pathways relatively depleted in NGM3 microbiota.

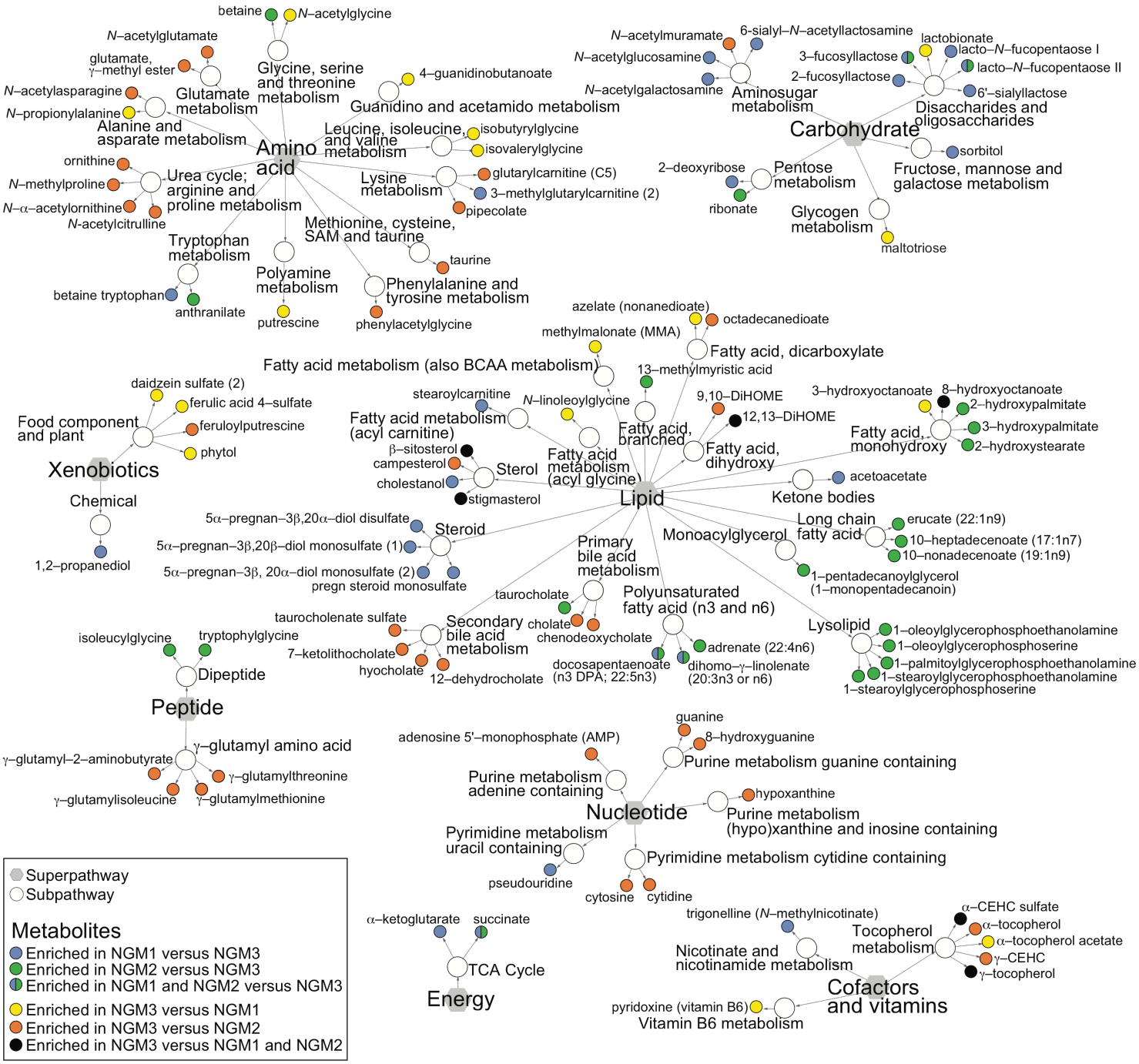

Fig. 6. Inter–NGMS comparisons reveal distinct programs of metabolism in the neonatal gut associated with PM–atopy development. Comparative UPLC–MS/MS–based metabolic profiling of neonatal representative feces from each of the three NGMs indicates that the lower–risk NGM1 (*n* = 10) and NGM2 (*n* = 10) subjects exhibit significant differences in metabolite relative concentration compared to high–risk NGM3 (*n* = 8; Welch's *t*–test; $P < 0.05$).

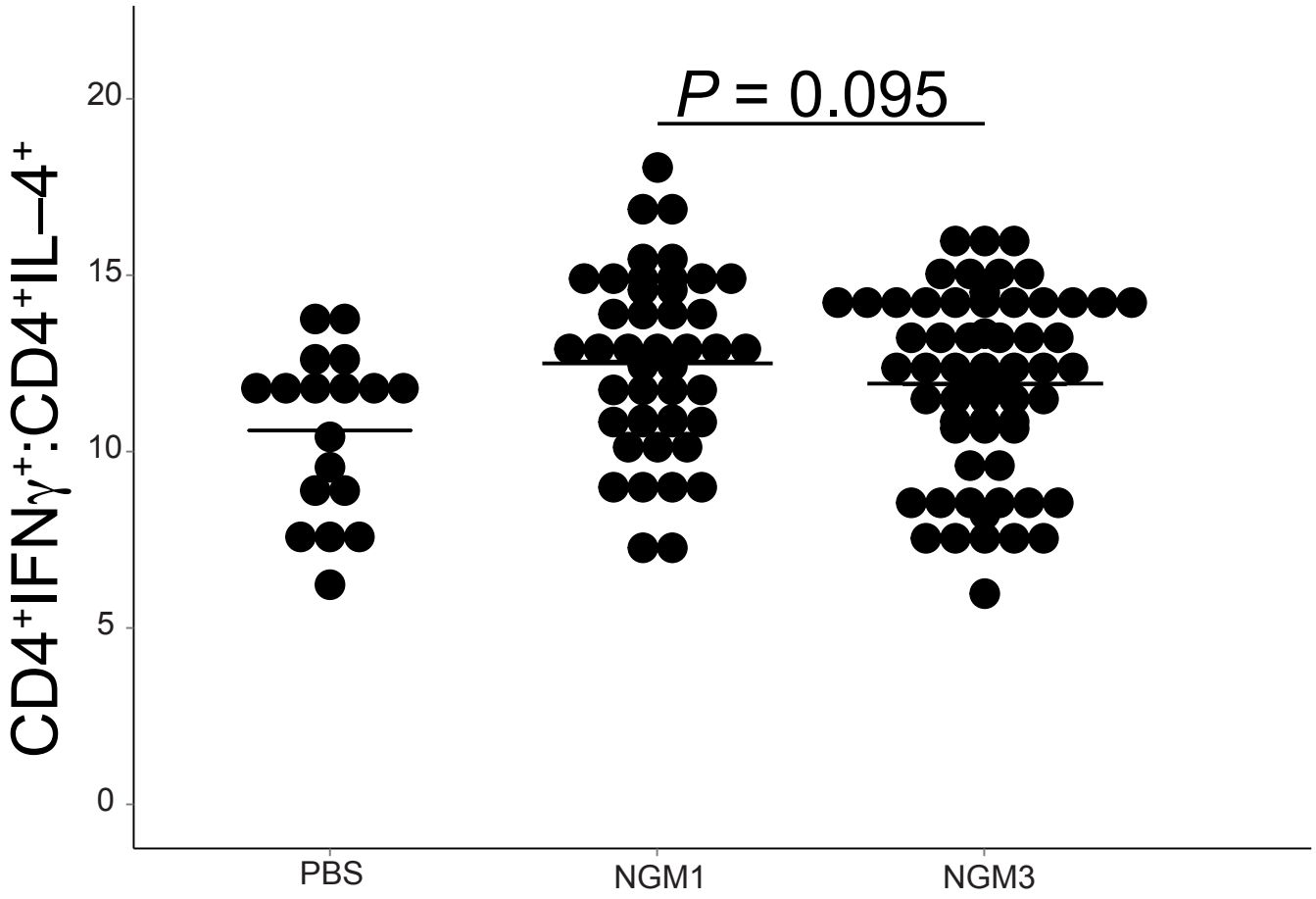

Fig. 7. Sterile fecal water from NGM3 induces a $CD4^+IL–4^+$ cell skew. Dendritic cells and autologously purified naïve $CD4^+$ cells from serum of two healthy adult donors (biological replicates), were incubated with sterile fecal water from NGM1 (*n* = 7; three biological replicates per sample) or NGM3 (*n* = 5; three biological replicates per sample) participants. NGM3 fecal water induces a trend toward a $CD4^+IL–4^+$ cell skew compared to NGM1 (LME; *P* = 0.095).

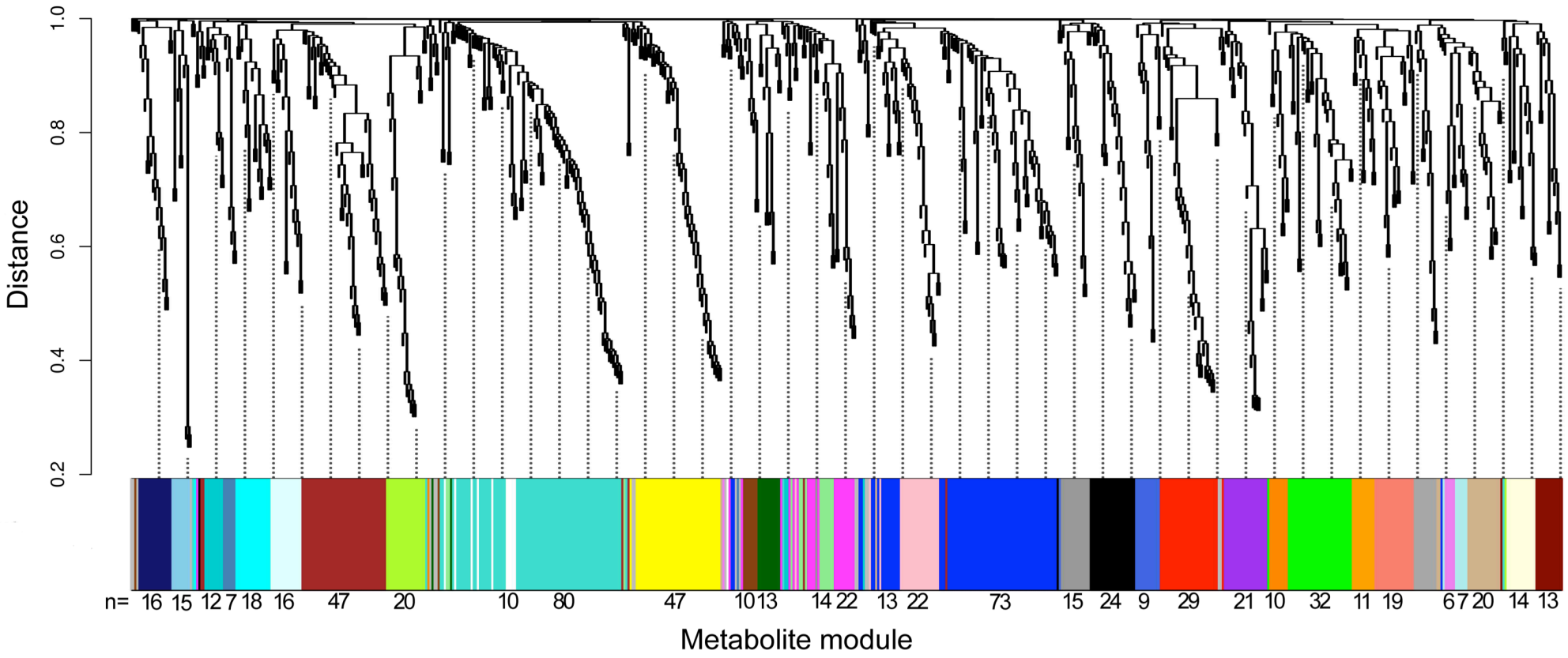

Fig. 8. Weighted correlation network analysis identifies modules of co-associated fecal metabolites detected using UPLC–MS/MS profiling. Pearson's correlation was used to determine inter–metabolite relationships. Each module, represented by a distinct color, corresponds to a group of positively co-associated metabolites (minimum five metabolites per module). Number of metabolites in each module is provided below each; grey bars represent unassigned biochemicals.

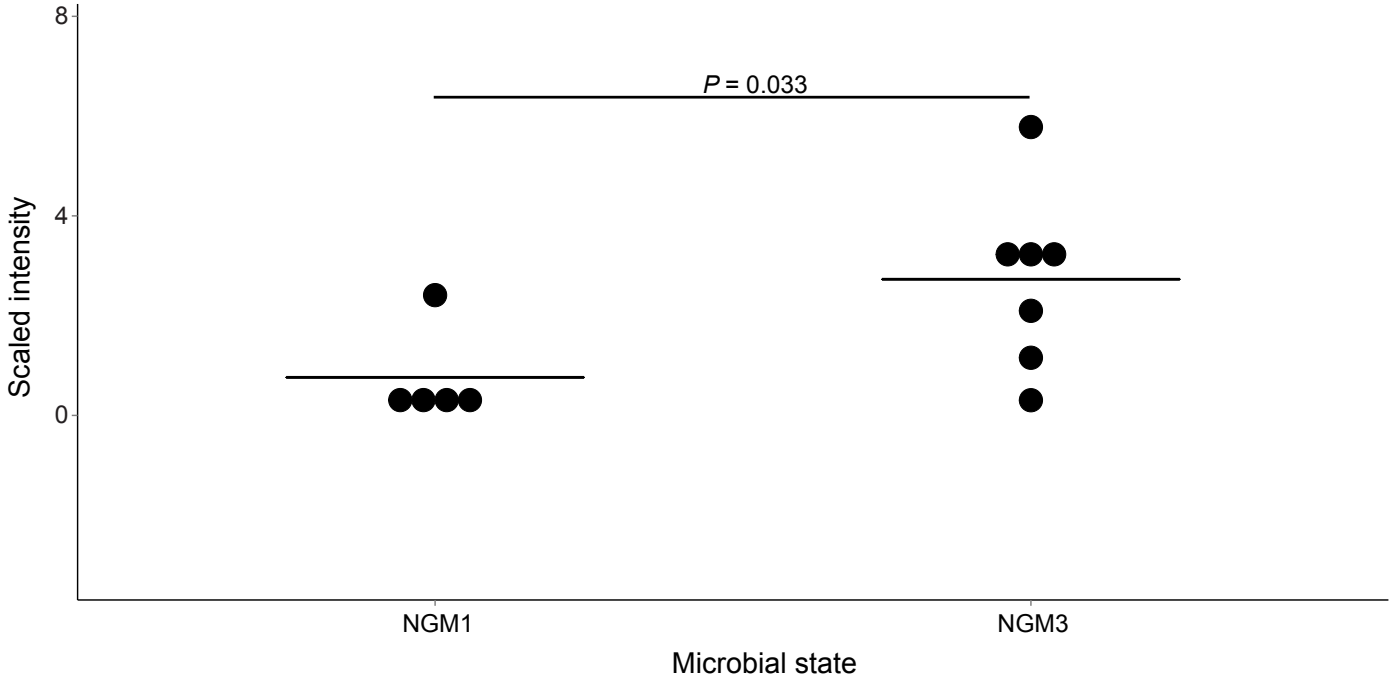

Fig. 9. Dihydroxy fatty acid 12, 13 DiHOME concentration is significantly increased in NGM3 sample subset used for *ex vivo* assays. Using the subset of samples employed in the *ex vivo* DC–T–cell assay and based on metabolite scaled intensity data obtained from UPLC–MS/MS data, 12, 13 DiHOME is significantly increased in relative concentration in NGM3 (n = 7) compared to NGM1 (n = 5) samples (Welch's t–test; P = 0.033).

Supplementary tables

Table 1. Allergens used to determine PM atopy status of participants in this study. Mean and median of allergen–specific IgE (IU ml^{-1}) is provided for each.

Allergen	*n*	Mean (SD)	Median	[Min, Max]
Alternaria	292	0.21 (0.86)	0.05	[0.05, 12.8]
German cockroach (Bla g 2)	296	0.22 (1.09)	0.05	[0.05, 14.7]
Dog (Can f 1)	295	0.2 (0.65)	0.05	[0.05, 6.22]
House dust mite (Der f 1)	295	0.18 (0.72)	0.05	[0.05, 7.55]
Egg	298	1.48 (10.64)	0.05	[0.05, 170]
Cat (Fel d 1)	295	0.23 (1.27)	0.05	[0.05, 14.8]
Milk	296	0.79 (3.65)	0.05	[0.05, 59.0]
Peanut	291	2.88 (34.07)	0.05	[0.05, 572]
Common ragweed	292	0.08 (0.12)	0.05	[0.05, 1.08]
Timothy grass	296	0.09 (0.23)	0.05	[0.05, 2.22]

Table 2. **Risk ratio of IGMs (infants > 6 months old) developing atopy or having parental report of doctor's diagnosis of asthma.** Risk ratios were calculated based on log–binomial regression.

	DMM community type		*RR* (95% CI)	
	IGM1 (*n* = 89)	IGM2 (*n* = 79)	IGM2 versus IGM1	*P*–value
Atopy (PM)	21 (23.6%)	19 (24.1%)	1.02 (0.59,1.75)	0.94
Parental report of doctor diagnosed asthma	15 (19.2%)	7 (9.7%)	0.51 (0.22, 1.17)	0.11
Atopy (IgE > 0.35 IU ml^{-1})	49 (55.1%)	38 (48.1%)	0.87 (0.65, 1.17)	0.37

Table 3: Association between early life factors and IGMs.

Table 4: Association between early life factors and NGMs.

Table 5. Factors tested for possible confounding effect on the risk of developing PM atopy for NGM.

Table 6. Bacterial taxa exhibiting significantly increased relative abundance in low–risk NGM1 versus the high–risk NGM3 neonatal gut microbiota.

Table 7. Bacterial taxa exhibiting significantly increased relative abundance in low–risk NGM2 versus the high–risk NGM3 neonatal gut microbiota.

Table 8. Fungal taxa exhibiting significantly increased relative abundance in low–risk NGM1 versus high–risk NGM3 neonatal gut microbiota. Significant difference in relative abundance was determined using a zero–inflated negative binomial regression model, $q < 0.20$. White background indicates taxa enriched in NGM1 compared to NGM3, gray background indicates taxa enriched in NGM3 compared to NGM1.

OTU	NGM1–NGM3	*q*–value	Phylum	Order	Family	Genus
1	7207.18	7.1E–11	Ascomycota	Saccharomycetales	Unclassified	Unclassified
17	803.31	6.7E–03	Basidiomycota	Malasseziales	Incertae sedis	*Malassezia*
430	188.46	6.7E–05	Basidiomycota	Malasseziales	Incertae sedis	*Malassezia*
25	150.36	6.2E–127	Unclassified	Unclassified	Unclassified	Unclassified
2188	62.78	3.2E–02	Ascomycota	Capnodiales	Davidiellaceae	Unclassified
997	42.42	1.7E–03	Ascomycota	Eurotiales	Trichocomaceae	Unclassified
109	22.28	2.2E–49	Unclassified	Unclassified	Unclassified	Unclassified
102	8.80	3.8E–11	Basidiomycota	Trichosporonales	Trichosporonaceae	Unclassified
228	4.21	1.0E–05	Ascomycota	Chaetothyriales	Unclassified	Unclassified
2111	2.72	1.0E–01	Unclassified	Unclassified	Unclassified	Unclassified
318	2.01	1.8E–01	Basidiomycota	Malasseziales	Incertae sedis	*Malassezia*
172	0.22	2.6E–02	Ascomycota	Saccharomycetales	Incertae sedis	*Candida*
2344	–919.11	6.2E–04	Basidiomycota	Sporidiobolales	Incertae sedis	*Rhodotorula*
84	–364.58	1.8E–29	Ascomycota	Hypocreales	Nectriaceae	Unclassified
145	–171.96	2.6E–14	Basidiomycota	Polyporales	Phanerochaetaceae	*Phanerochaete*
94	–70.57	3.5E–17	Ascomycota	Pleosporales	Pleosporaceae	Unclassified
23	–52.81	3.1E–07	Ascomycota	Unclassified	Unclassified	Unclassified
107	–35.81	1.4E–06	Ascomycota	Saccharomycetales	Incertae sedis	Unclassified
29	–35.64	1.4E–06	Ascomycota	Saccharomycetales	Incertae sedis	*Candida*
69	–32.10	1.9E–12	Ascomycota	Saccharomycetales	Incertae sedis	*Cyberlindnera*
62	–31.93	1.2E–06	Ascomycota	Saccharomycetales	Debaryomycetaceae	*Meyerozyma*
273	–31.29	3.9E–03	Ascomycota	Unclassified	Unclassified	Unclassified
260	–10.47	3.0E–43	Ascomycota	Pleosporales	Incertae sedis	Unclassified
2018	–9.79	2.7E–19	Ascomycota	Saccharomycetales	Saccharomycetaceae	*Saccharomyces*
637	–7.64	5.1E–03	Basidiomycota	Sporidiobolales	Incertae sedis	*Rhodotorula*
367	–7.60	2.4E–03	Basidiomycota	Unclassified	Unclassified	Unclassified
473	–2.96	1.0E–03	Ascomycota	Pleosporales	Cucurbitariaceae	*Pyrenochaetopsis*
745	–1.31	4.7E–03	Ascomycota	Saccharomycetales	Incertae sedis	*Candida*
1971	–0.91	1.2E–02	Ascomycota	Saccharomycetales	Saccharomycetaceae	*Saccharomyces*

Table 9. Fungal taxa exhibiting significantly increased relative abundance in low–risk NGM2 versus high–risk NGM3 neonatal gut microbiota. Significant difference in relative abundance was determined using zero–inflated negative binomial regression model, $q < 0.20$. White background indicates taxa enriched in NGM2 compared to NGM3, gray background indicates taxa enriched in NGM3 compared to NGM2.

OTU	NGM2–NGM3	*q*–value	Phylum	Order	Family	Genus
17	1462.83	2.5E–05	Basidiomycota	Malasseziales	Incertae sedis	*Malassezia*
430	180.12	5.5E–05	Basidiomycota	Malasseziales	Incertae sedis	*Malassezia*
2113	88.31	1.9E–01	Basidiomycota	Malasseziales	Incertae sedis	*Malassezia*
656	80.83	4.9E–02	Basidiomycota	Malasseziales	Incertae sedis	*Malassezia*
2188	64.25	1.0E–01	Ascomycota	Capnodiales	Davidiellaceae	Unclassified
997	22.41	1.5E–02	Ascomycota	Eurotiales	Trichocomaceae	Unclassified
171	19.59	1.1E–140	Unclassified	Unclassified	Unclassified	Unclassified
102	12.02	1.6E–56	Basidiomycota	Trichosporonales	Trichosporonaceae	*Trichosporon*
2252	7.46	2.0E–01	Basidiomycota	Malasseziales	Incertae sedis	*Malassezia*
165	6.18	1.9E–01	Ascomycota	Eurotiales	Trichocomaceae	Unclassified
2111	3.01	1.8E–02	Unclassified	Unclassified	Unclassified	Unclassified
878	0.78	4.1E–05	Unclassified	Unclassified	Unclassified	Unclassified
2277	0.76	3.8E–03	Unclassified	Unclassified	Unclassified	Unclassified
1884	0.45	8.2E–02	Unclassified	Unclassified	Unclassified	Unclassified
1309	0.08	8.2E–02	Unclassified	Unclassified	Unclassified	Unclassified
2344	–920.65	3.3E–04	Basidiomycota	Sporidiobolales	Incertae sedis	*Rhodotorula*
84	–364.28	5.9E–70	Ascomycota	Hypocreales	Nectriaceae	Unclassified
963	–178.30	1.1E–01	Ascomycota	Saccharomycetales	Incertae sedis	*Candida*
28	–164.94	1.1E–15	Ascomycota	Saccharomycetales	Incertae sedis	*Debaryomyces*
23	–52.77	1.6E–04	Ascomycota	Unclassified	Unclassified	Unclassified
107	–35.06	2.5E–22	Ascomycota	Saccharomycetales	Incertae sedis	Unclassified
62	–29.90	1.7E–03	Ascomycota	Saccharomycetales	Debaryomycetaceae	*Meyerozyma*
187	–20.35	1.0E–12	Ascomycota	Trichosphaeriales	Incertae sedis	*Nigrospora*
260	–10.68	6.4E–20	Ascomycota	Pleosporales	Incertae sedis	Unclassified
2018	–9.91	5.1E–13	Ascomycota	Saccharomycetales	Saccharomycetaceae	*Saccharomyces*
473	–2.94	1.7E–03	Ascomycota	Pleosporales	Cucurbitariaceae	*Pyrenochaetopsis*
1505	–2.15	1.7E–13	Unclassified	Unclassified	Unclassified	Unclassified
1944	–1.80	1.8E–02	Unclassified	Unclassified	Unclassified	Unclassified
745	–1.27	1.7E–02	Ascomycota	Saccharomycetales	Incertae sedis	*Candida*
885	–0.37	1.8E–02	Unclassified	Unclassified	Unclassified	Unclassified

Table 10: Procrustes analyses of 16S rRNA phylogeny, PICRUSt and metabolomics datasets. Results from Procrustes analyses indicate that 16S rRNA phylogeny, PICRUSt and metabolomics data is highly and significantly correlated.

Comparison	r^* (M^2)	*P*–value
16S rRNA versus PICRUSt	0.72 (0.48)	< 0.001
16S rRNA versus Metabolomics	0.87 (0.24)	< 0.001
PICRUSt versus Metabolomics	0.66 (0.56)	0.010

**r* = correlation between data sources. Unweighted UniFrac distance used for 16S rRNA; Canberra distance used for PICRUSt and Metabolomics.

Table 11. Metabolites significantly enriched in low–risk NGM1 versus high–risk NGM3 neonatal gut microbiota.

Table 12. Metabolites significantly enriched in low–risk NGM2 versus high–risk NGM3 neonatal gut microbiota.

Supplementary information

Gut microbiota–state validation

In order to assess the validity of our DMM modeling, the published 16S rRNA data of Arrieta *et al.*[1] was used (*n* = 319 independent fecal samples collected at approximately 3–12 months of age). The specific age of each participant was unavailable and the youngest participants in this cohort were 3 months of age, substantially older than neonates in the WHEALS cohort. Hence the dataset could not be segregated into samples that were > or < 6 months of age, as had been performed for our WHEALS cohort, which limited our capacity to identify neonatal microbiota states associated with subsequent childhood atopy and asthma outcomes. Because of the age range of the CHILD cohort, we applied both our NGM and IGM model parameters to the entire data set. A better model fit (i.e., smaller laplace approximation to the negative log model evidence) was obtained when the CHILD data was fit to the NGM model compared to the IGM model (model fit: 32,502 versus 174,610, respectively) and a two–group solution represented the best fit. Group 1 (G1) included 221 (69%) participants and group 2 (G2) 98 (31%). The posterior probabilities were on average higher for G1 compared to G2 (0.98 vs. 0.95, respectively). Consistent with our findings, CHILD participants assigned to G1 were typically defined by high Bifidobacteriaceae relative abundance (average relative abundance (aRA): 75%). G2 participants were characterized by Lachnospiraceae (aRA: 39%), Clostridiaceae (aRA: 29%), and Ruminococcaceae (aRA: 12%), more reflective of the IGM2 cluster identified in our cohort.

Code availability

The following script may be used to calculate a representative multiply rarefied OTU table from an unrarefied OTU table, an alterative to single rarefied tables that stabilizes the effect of random sampling and results in an OTU table that is more representative of community composition. Multiple single–rarefied OTU tables are calculated for each sample, and the distance between the subject–specific rarefied vectors calculated. The rarefied vector that is the minimum average (or median) distance from itself to all other rarefied vectors is considered the most representative for that subject and used to represent community composition for that sample in the resulting multiply–rarified OTU table.

```
library(vegan)
library(GUNifFrac)

##Parameters
# specify the raw OTU count table, with samples = rows, taxa = columns
# rawtab = otu_tab_t
```

```
# specify the depth you would like to rarefy your tables to the default is to just use the minimum sequencing
#depth raredepth = min(rowSums(rawtab))

# specify the number of rarefied tables you would like to generate to calculate your representatiave rarefied
#table from ntables = 100

# specify the distance measure to use to calculate distance between rarefied data sets, for each subject
#can be any of the methods available in the vegdist function of vegan distmethod = "euclidean"

# specify the method to summarize across distances if mean distance, then summarymeasure = mean
#if median distance, then summarymeasure = median
# summarymeasure = mean

# specify the seed start for the rarefied tables
# for each subsequent table, 1 will be added that the previous seed
# for reproducibility, always save your seedstart value (or just use the default for simplicity).
# seedstart = 500

# specify if you want progress updates to be printed
# verbose = TRUE

### returns a representative rarefied OTU table of class matrix.

##functions
reprare <– function(rawtab = otu_tab_t, raredepth = min(rowSums(otu_tab_t)), ntables = 100, distmethod =
euclidean",
summarymeasure=mean, seedstart = 500, verbose = TRUE) {
raretabs = list()
for (z in 1:ntables) {
if (verbose == TRUE) {
print(paste("calculating rarefied table number", z, sep = " "))
}
set.seed(seedstart + z)
raretabs[[z]] = Rarefy(rawtab, depth = raredepth)[[1]]
```

```
}
raretabsa = array(unlist(raretabs), dim = c(nrow(raretabs[[z]]), ncol(rawtab), ntables))
final_tab = c()
for (y in 1:nrow(raretabs[[z]])) {
if (verbose == TRUE) {
print(paste("determining rep rarefied vector for subject number", y, sep = " "))
}
distmat = as.matrix(vegdist(t(raretabsa[y,,]), method = distmethod)) # distance across reps for subject y
distsummary = apply(distmat, 2, summarymeasure)
whichbestrep = which(distsummary == min(distsummary))[1]  # the best rep is the one with the minimum
average/median distance to all other reps. (in case of ties, just select the first)
bestrep = raretabsa[y,,whichbestrep] # select that rep only for subject y
final_tab = rbind(final_tab, bestrep) # build that rep for subject y into final table
}
rownames(final_tab) = rownames(raretabs[[z]])
colnames(final_tab) = colnames(rawtab)
return(final_tab)
}

###### example runs of the function:  ######
### dummy data set for example ###
ntaxa = 200
nsubj = 50
set.seed(444)

dummyOTU <– matrix(sample(0:500, ntaxa*nsubj, prob = c(0.7,0.1,0.1,rep(0.1/498, 498)), replace = TRUE),
ncol = ntaxa)
colnames(dummyOTU) = paste("OTU", 1:ntaxa, sep = "")
rownames(dummyOTU) = paste("subj", 1:nsubj, sep = "")
sort(rowSums(dummyOTU)) # sequencing depth is uneven

# specify the minimum depth
repraretable = reprare(rawtab = dummyOTU, raredepth = min(rowSums(dummyOTU)), ntables = 100,
distmethod = "euclidean",
```

```
summarymeasure = mean, seedstart = 500, verbose = TRUE)
dim(repraretable)
sort(rowSums(repraretable)) # sequencing depth is now even

# specify a depth other than the minimum
repraretable = reprare(rawtab = dummyOTU, raredepth = 3380, ntables = 100, distmethod = "euclidean",
summarymeasure = mean, seedstart = 500, verbose = TRUE)
dim(repraretable) # subjects with less than the minimum are no longer in the table
sort(rowSums(repraretable)) # sequencing depth is now even
```