Supplementary Materials for

Fecal microbiome and metabolome differ in healthy and food-allergic twins

Authors: Riyue Bao[1,2,3], Lauren A. Hesser[4], Ziyuan He[5], Xiaoying Zhou[5], Kari C. Nadeau[5,6,7]‡, Cathryn R. Nagler[4,8]‡*

Affiliations:

[1]Department of Pediatrics, University of Chicago, Chicago, IL, USA.

[2]UPMC Hillman Cancer Center, Pittsburgh, PA, USA.

[3]Department of Medicine, University of Pittsburgh, Pittsburgh, PA, USA.

[4]Pritzker School of Molecular Engineering, The University of Chicago, Chicago, IL, USA.

[5]Sean N. Parker Center for Allergy and Asthma Research at Stanford University, Stanford University, Stanford, CA, USA.

[6]Division of Pulmonary and Critical Care Medicine, Stanford University, Stanford, CA, USA.

[7]Division of Allergy, Immunology and Rheumatology, Department of Medicine, Stanford University, Stanford, CA, USA.

[8]Department of Pathology, The University of Chicago, Chicago, IL, USA.

‡Co-senior authors

*To whom correspondence should be addressed:

Cathryn Nagler, Ph.D.

Bunning Food Allergy Professor

Biological Sciences Division and

Pritzker School of Molecular Engineering

The University of Chicago

Jules F. Knapp Medical Research Building

924 East 57th Street, R410

Chicago, IL 60637

Phone: 773-702-6317

Fax: 773-702-3993

Email: cnagler1@uchicago.edu

This PDF includes:

Fig. S11. Correlation between the 64 OTUs from Fig. 3 and the 97 metabolites from Fig. 5a and 5b.

Fig. S12. Distribution of metabolite Spearman's correlation coefficient between healthy-abundant OTU clusters 1 to 3 for each metabolite group from Fig. 7.

Fig. S13. Quantitative PCR (qPCR) validation of *Phascolarctobacterium faecium* discovered by the 16S sequencing platform, shown in discordant twin pairs (10 pairs, n=20).

Fig. S14. qPCR validation of *Ruminococcus bromii* discovered by the 16S sequencing platform, shown in discordant twin pairs (10 pairs, n=20).

Tables S1 to S13 are provided as Supplementary Data File in Excel Spreadsheet format.

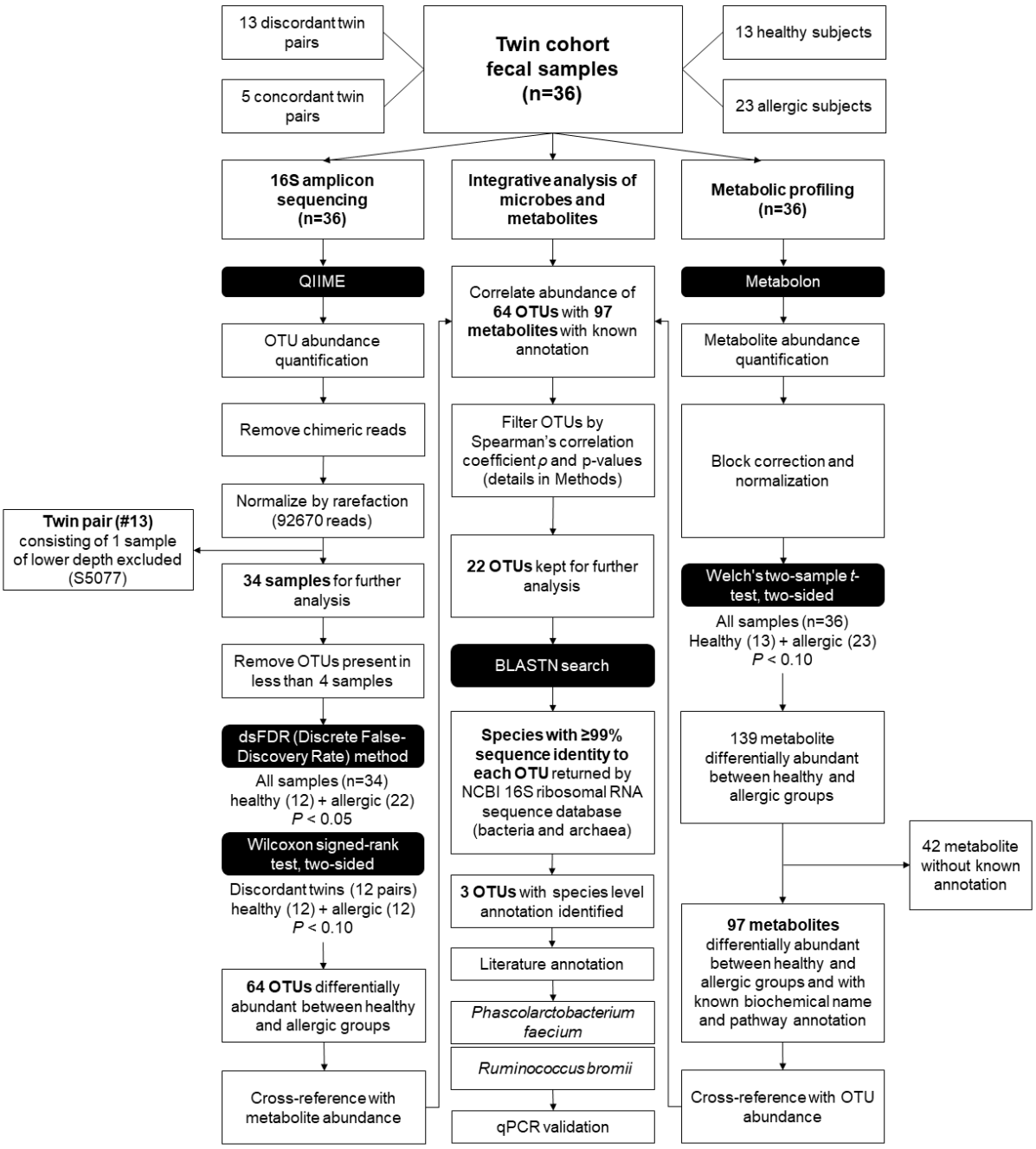

Fig. S1. Overview of the analysis workflow on the microbial 16S sequencing data, metabolite profiling data, and integration of the two types of data.

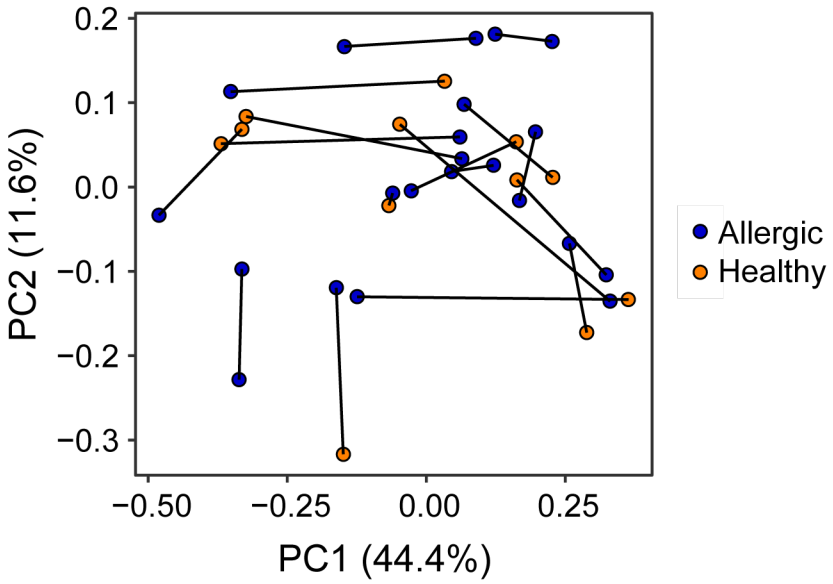

Fig. S2. β-diversity Principal Coordinates Analysis (PCoA) of twin fecal microbial communities with weighted UniFrac measure. Shown is a plot of the first two principal coordinate axes (PC1 and PC2) explaining 44.4% and 11.6% of the total variance among 34 samples from the healthy and allergic twins. Each dot represents one sample. Line connects samples from the same twin pair. PERMANOVA was used to test the diversity differences between healthy and allergic groups (P=0.82).

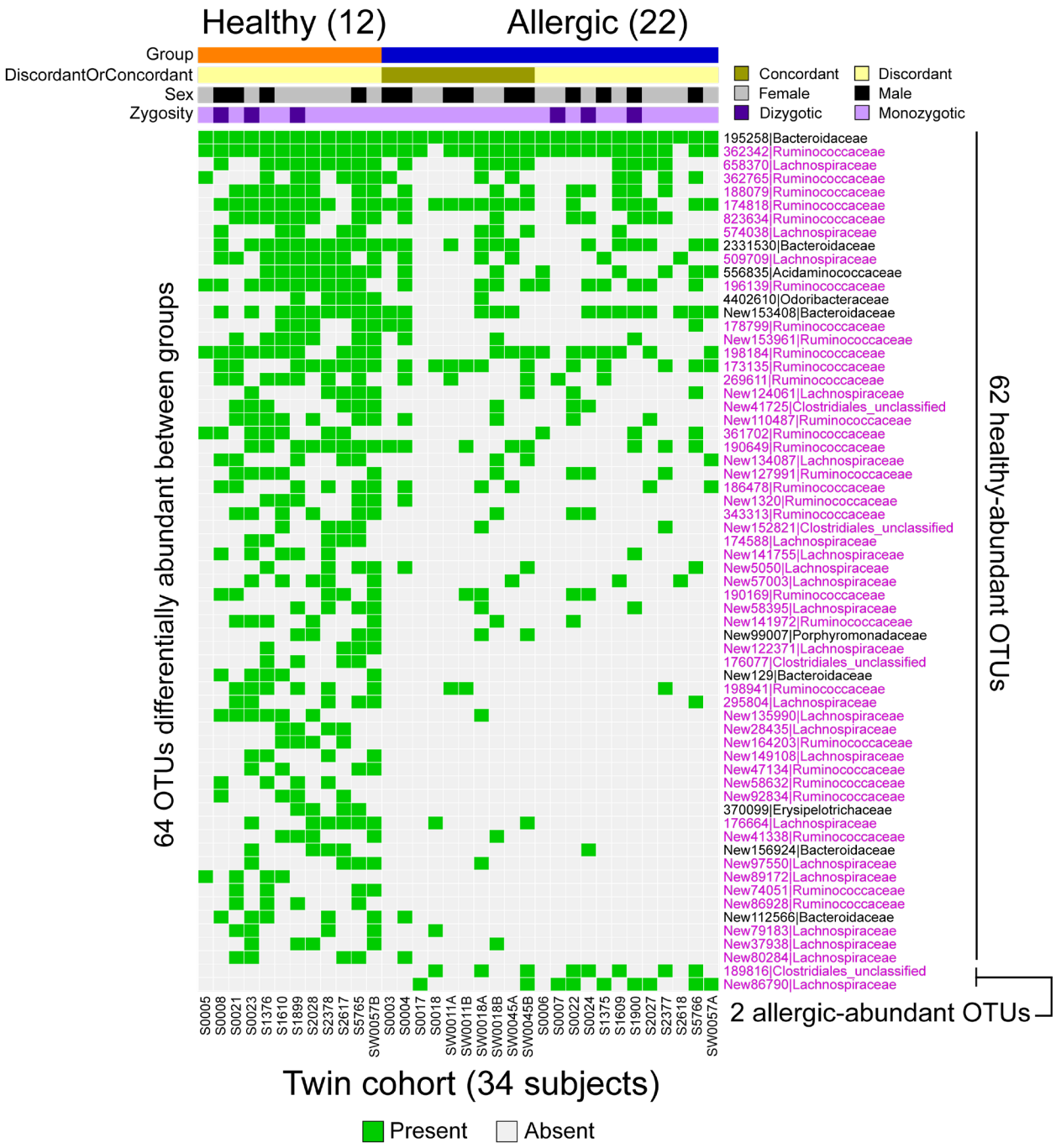

Fig. S3. Binary heatmap of the 64 OTUs differentially abundant between healthy and allergic groups. Green indicates the presence of an OTU in a sample, and light grey indicates absence. The 64 OTUs were from **Fig. 3**. Out of 64 OTUs, 62 are more abundant in the healthy group, and 2 are more abundant in the allergic group. OTU IDs are shown on the row in the format of "OTU_ID|Family", and those annotated with family Lachnospiraceae,

Ruminococcaceae, or Clostridiales_unclassified are highlighted in pink. Sample IDs are shown on the column, with annotation bars above the heatmap indicating concordant/discordant twin members, sex, and zygosity.

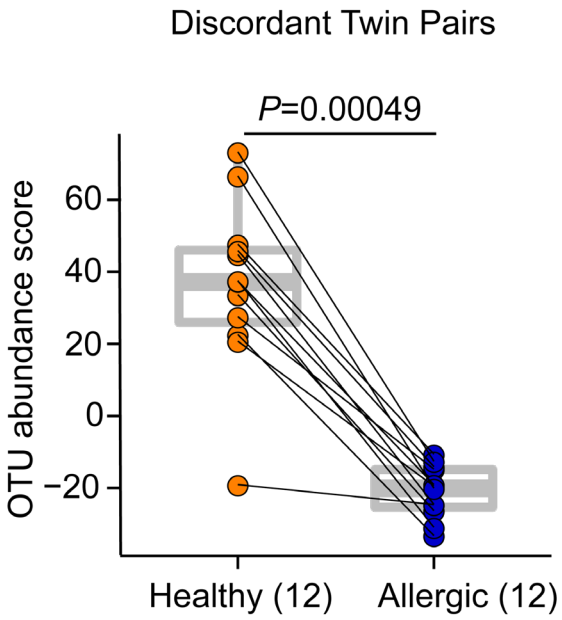

Fig. S4. The aggregated OTU abundance score is significantly higher in healthy relative to allergic group in the discordant twin pairs (12 pairs, *n*=24). The result of all 34 samples is shown in **Fig. 4b**. Each dot denotes one sample. The bounds of the boxes represent the 25th and 75th percentiles, the horizontal center line indicates the median, and the whiskers extend to data points within a maximum of 1.5 times the interquartile range (IQR). Two-sided Wilcoxon signed-rank test was used to compute the p-value.

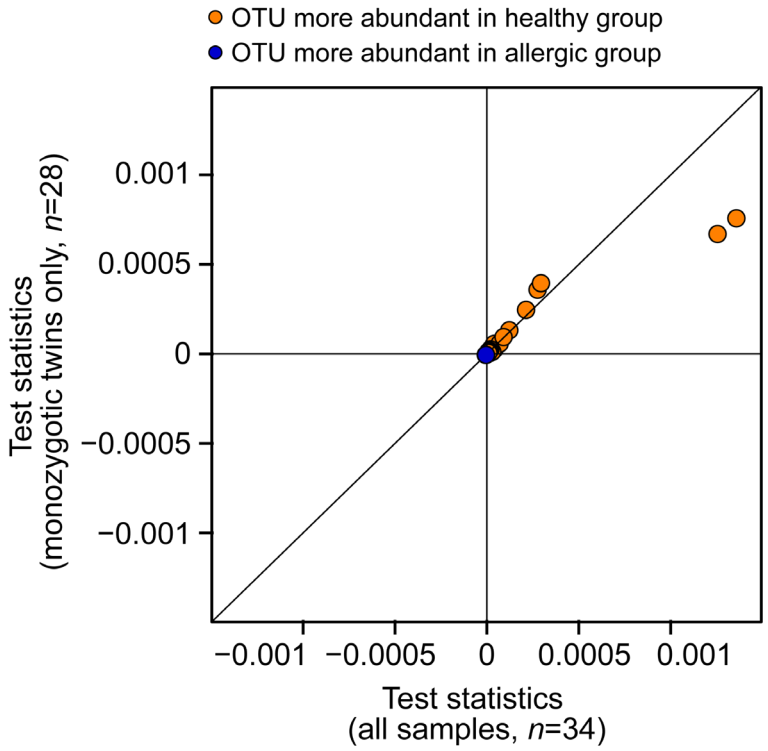

Fig. S5. Test statistics of the differentially abundant OTUs from all samples (*n*=34) is correlated with that computed from monozygotic twins only (*n*=28) comparing healthy and allergic groups. For each of the 64 OTUs differentially abundant between healthy and allergic groups using all samples, 51 are present in at least 4 out of the 28 monozygotic twin samples, hence were included for re-computing the DS-FDR permutation test statistics using samples from monozygotic twins only, and are shown on the figure.

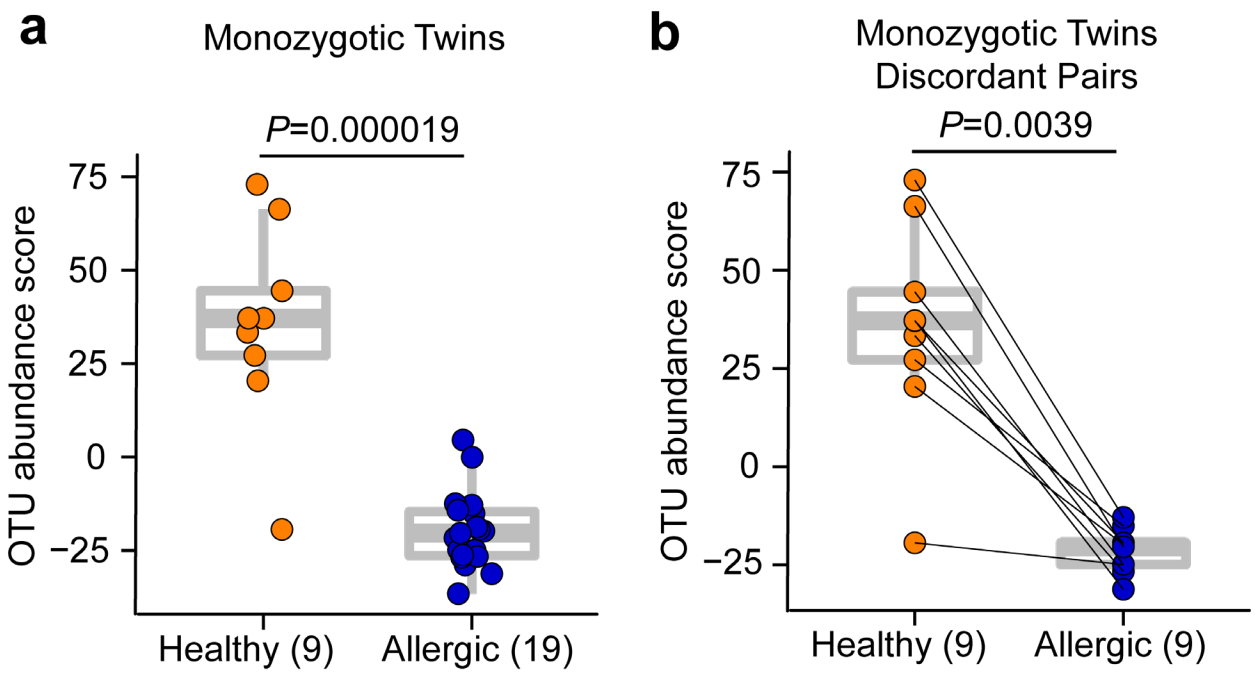

Fig. S6. The aggregated OTU abundance score remains significant in healthy relative to allergic group in monozygotic twins (*n*=28). (**a**) 28 samples from 14 pairs of monozygotic twins. (**b**) 18 samples from 9 pairs of monozygotic twins that are discordant. Each dot denotes one sample. The bounds of the boxes represent the 25th and 75th percentiles, the horizontal center line indicates the median, and the whiskers extend to data points within a maximum of 1.5 times the IQR. DS-FDR was used in **a**, two-sided Wilcoxon rank-sum test was used in **b**.

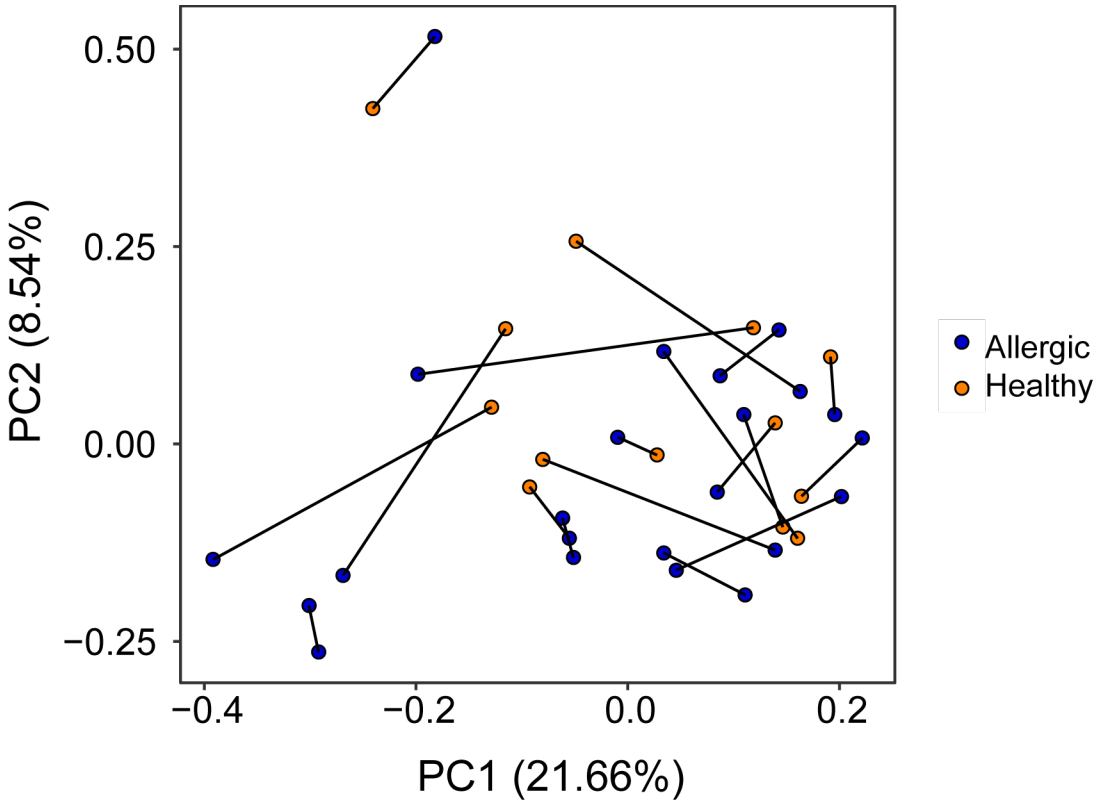

Fig. S7. Principle Component Analysis (PCA) of twin fecal metabolites. Shown is a plot of the first two principal component axes (PC1, PC2) explaining 21.66% and 8.54% of the total variance among 36 samples from the healthy and allergic groups. The one sample (S5077) and the corresponding twin pair (#13) excluded from 16S analysis due to low sequencing depth was included for metabolite analysis. Each dot represents one sample. Line connects samples from the same twin pair. PCA was performed on normalized and log10-transformed quantification of 1,308 metabolites in total.

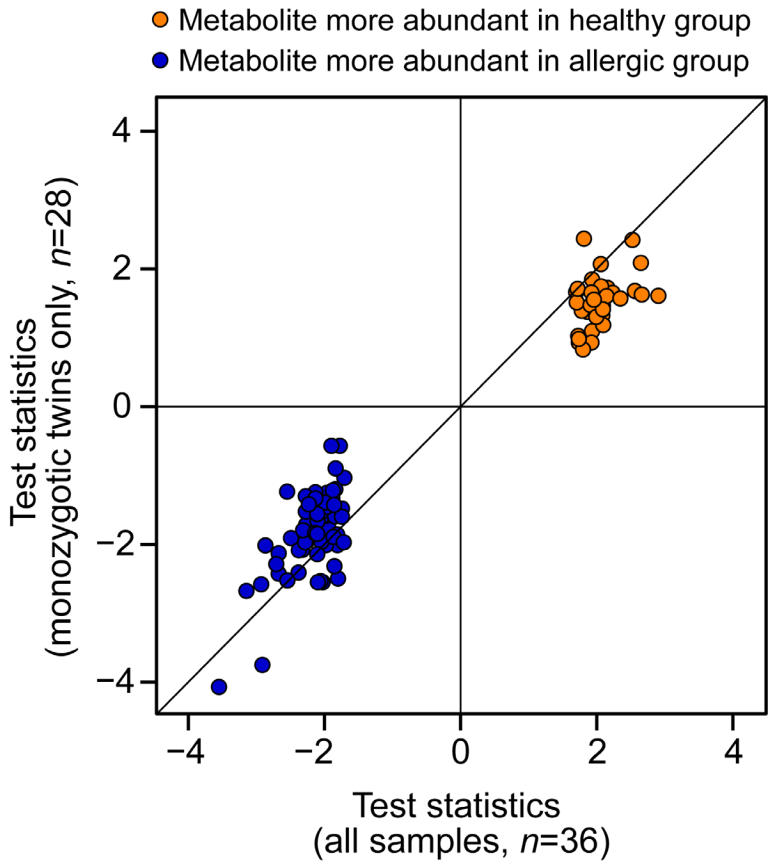

Fig. S8. Test statistics of the differentially abundant metabolites from all samples (*n*=36) is correlated with that computed from monozygotic twins only (*n*=28) comparing healthy and allergic groups. For each of the 97 metabolites differentially abundant between healthy and allergic groups using all samples, test statistics was re-computed using samples from monozygotic twins only, and are shown on the figure. Two-sided Wilcoxon signed-rank test was used.

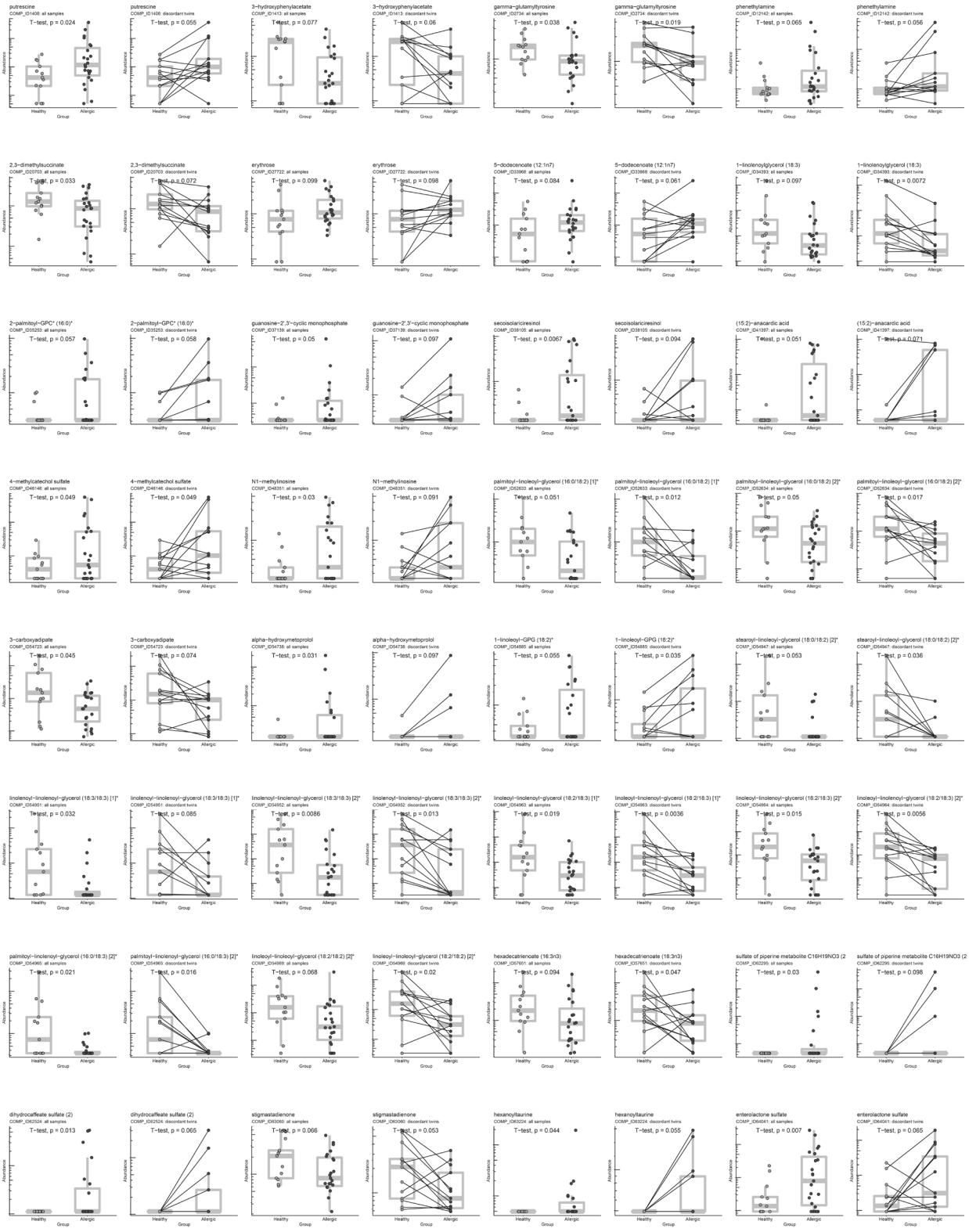

Fig. S9. 32 metabolites differentially abundant between healthy and allergic group in the discordant twin pairs (13 pairs, *n*=26). This is a subset of the 97 metabolites differentially

abundant between healthy and allergic groups across all 36 samples. The one sample (S5077) and the corresponding twin pair (#13) excluded from 16S analysis due to low sequencing depth was included for metabolite analysis, forming 13 discordant twin pairs. Four metabolites are shown per row. For each metabolite, two panels are shown: comparison between the two groups across all samples (left, n=36), and within discordant twin pairs only (right, 13 pairs, n=26), hence eight panels per row. *P*-values are shown in each panel. For comparison across all samples, two-sided Welch Two-Sample *t*-test was used; for comparison within discordant twin pairs only, two-sided paired *t*-test was used. All measure was normalized and log10-transformed before statistical tests (see **Methods**). The bounds of the boxes represent the 25th and 75th percentiles, the horizontal center line indicates the median, and the whiskers extend to data points within a maximum of 1.5 times the IQR.

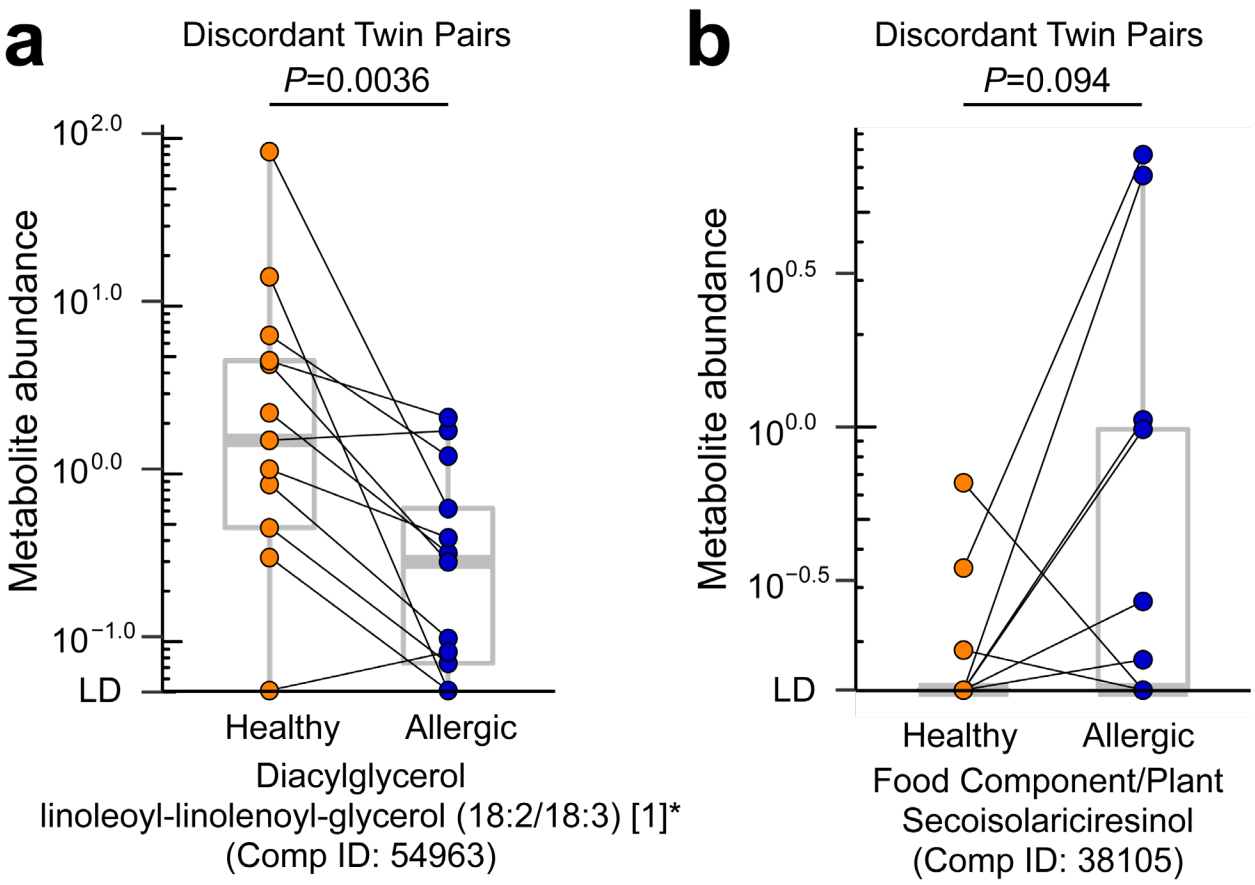

Fig. S10. Representative examples of metabolites significantly higher in healthy relative to allergic group in the discordant twin pairs, or vice versa (13 pairs, *n*=26). The two metabolites shown are from **Fig. 6b** and **6c**. Units shown on the y-axis in **a** and **b** represent the normalized raw area counts of UPLC MS/MS peaks, rescaled to set the median equal to 1.00 for each biochemical (see **Methods**). Each dot denotes one sample. The bounds of the boxes represent the 25th and 75th percentiles, the horizontal center line indicates the median, and the whiskers extend to data points within a maximum of 1.5 times the IQR. Two-sided paired *t*-test was used to compute the p-values.

Fig. S11. Correlation between the 64 OTUs from Fig. 3 and the 97 metabolites from Fig. 5a and 5b. Metabolites are shown on the row in the format of "COMP_ID|Biochemical_Name|Super_Pathway|Sub_Pathway", and OTU IDs are shown on the column in the format of "OTU_ID|Family". On the heatmap, between each OTU and each

metabolite, a positive correlation is shown in red, and a negative correlation is shown in blue. Spearman's correlation was used.

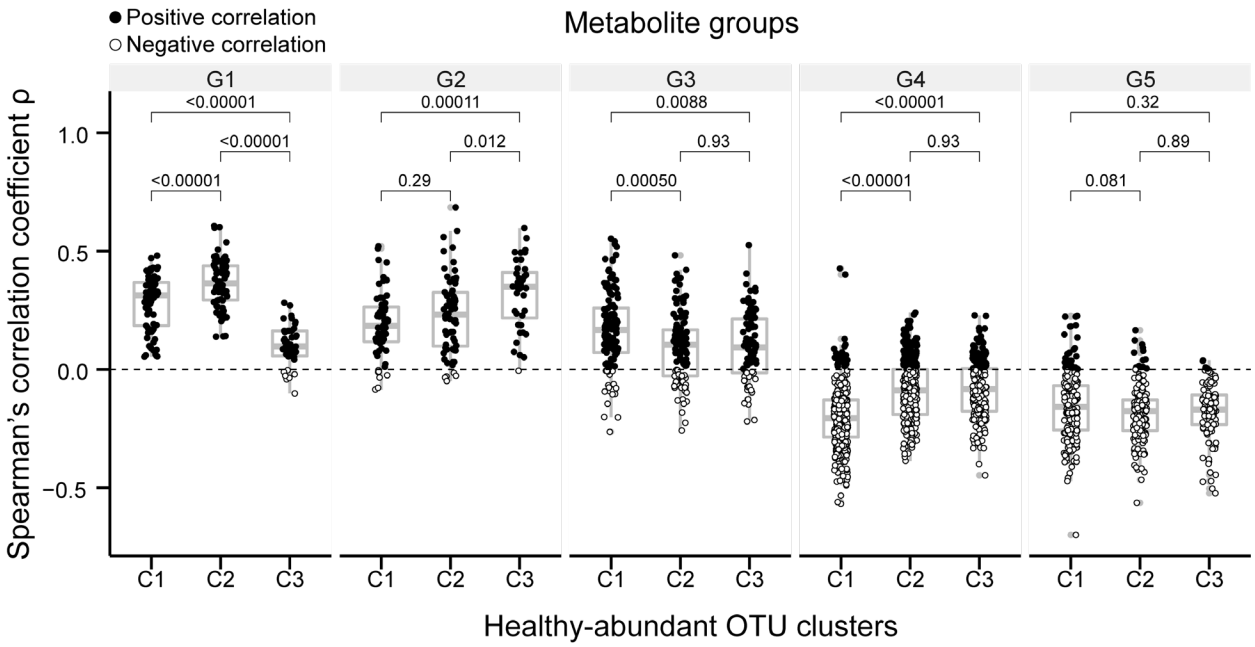

Fig. S12. Distribution of metabolite Spearman's correlation coefficient between healthy-abundant OTU clusters 1 to 3 for each metabolite group from Fig. 7. Spearman's correlation coefficient ρ from **Fig. 7** is shown on the y-axis. Healthy-abundant OTU clusters 1 to 3 from **Fig. 7** are shown on the x-axis. Pairwise comparison p-values are computed between C1/C2, C1/C3, and C2/C3 for each metabolite group. Tukey's honestly significant difference (HSD) test was used which controls false discovery rate for multiple comparisons. Each dot denotes one metabolite. The bounds of the boxes represent the 25th and 75th percentiles, the horizontal center line indicates the median, and the whiskers extend to data points within a maximum of 1.5 times the IQR.

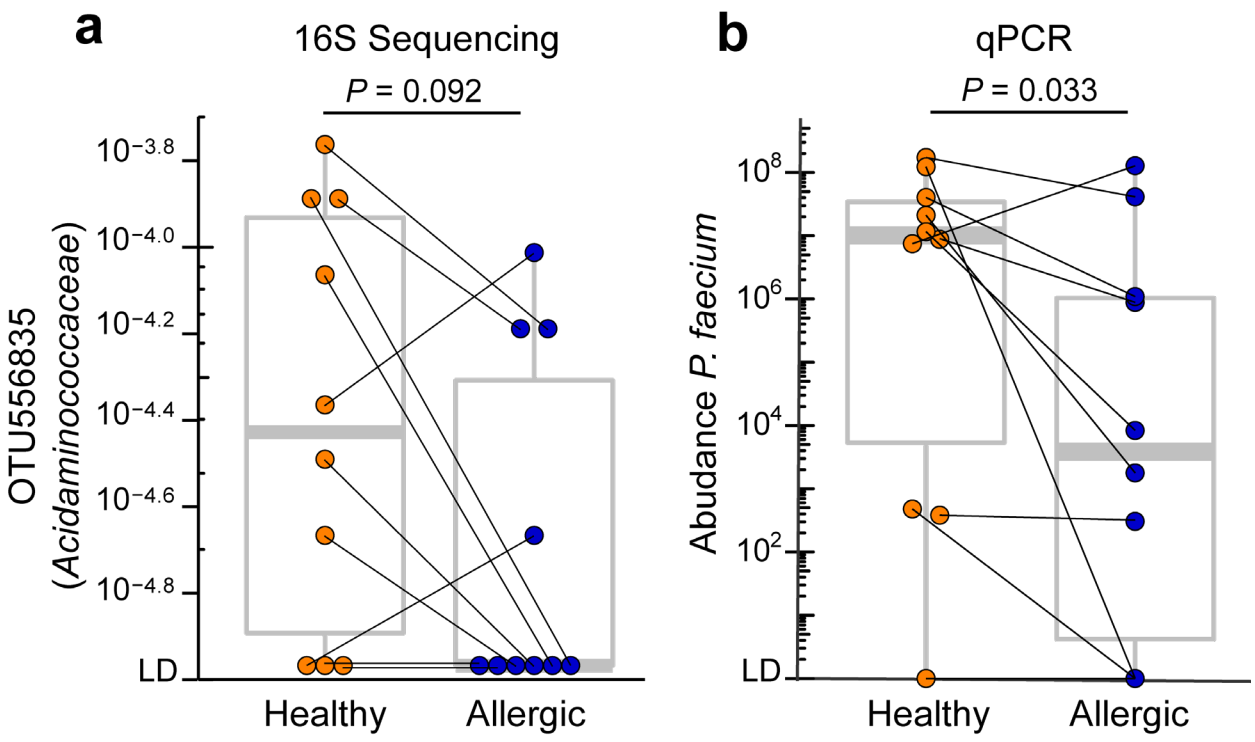

Fig. S13. Quantitative PCR (qPCR) validation of *Phascolarctobacterium faecium* discovered by 16S sequencing platform, shown in discordant twin pairs (10 pairs, *n*=20). 2 out of 12 discordant pairs did not have DNA materials left for qPCR validation, hence not included. The result of all samples is shown in **Fig. 8b** and **8c**. (**a**) OTU 556835 (family Acidaminococcaceae) is significantly more abundant in healthy compared to allergic group by 16S sequencing. This OTU was annotated as *Phascolarctobacterium faecium* at the species level with 99% sequence identity (NCBI accession ID NR_026111.1). P-value was re-calculated amongst the 10 twin pairs shown here from the 16S sequencing data, instead of 12 twin pairs total. (**b**) Quantitative PCR (qPCR) validates the abundance differences between healthy and allergic groups using *P. faecium*-specific primers. Units shown on the y-axis in represent 2^{-Ct} normalized to total 16S rRNA copies per gram of fecal material and multiplied by a constant ($1x10^{22}$) to bring all values above 1 (see **Methods**). Samples with abundance above the detection limit in both platforms are shown. Each dot denotes

one metabolite. The bounds of the boxes represent the 25th and 75th percentiles, the horizontal center line indicates the median, and the whiskers extend to data points within a maximum of 1.5 times the IQR. Two-sided Wilcoxon signed-rank test was used in **a** and **b**. qPCR data in **b** were log10 transformed before statistical testing.

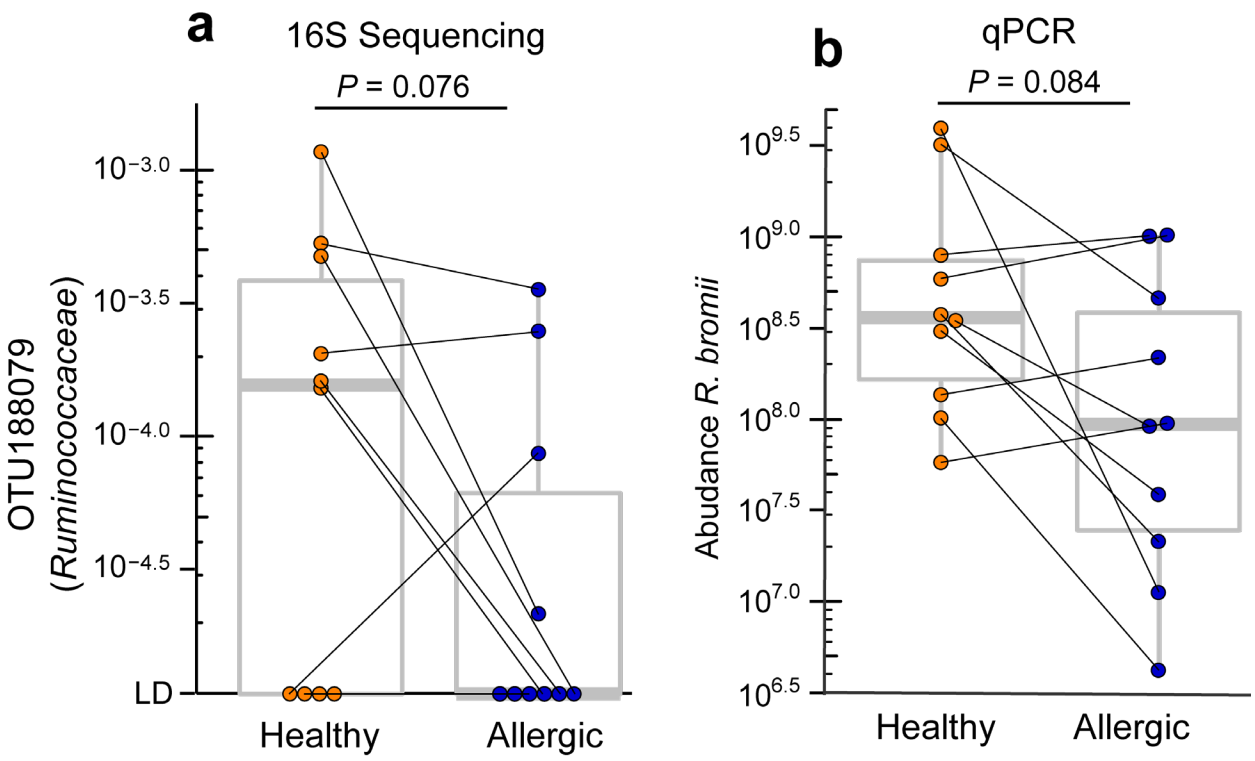

Fig. S14. Quantitative PCR (qPCR) validation of *Ruminococcus bromii* discovered by 16S sequencing platform, shown in discordant twin pairs (10 pairs, *n*=20). 2 out of 12 discordant pairs did not have DNA materials left for qPCR validation, hence not included. The result of all samples is shown in **Fig. 8d** and **8e**. (**a**) OTU188079 (family Ruminococcaceae) is significantly more abundant in healthy compared to allergic group by 16S sequencing. This OTU was annotated as *Ruminococcus bromii* at the species level with 99% sequence identity (NCBI accession ID NR_025930.1). P-value was re-calculated amongst the 10 twin pairs shown here from the 16S sequencing data, instead of 12 twin pairs total. (**b**) qPCR validates the abundance differences between healthy and allergic groups using *R. bromii*-specific primers. Units shown on the y-axis represent 2^{-Ct} normalized to total 16S rRNA copies per gram of fecal material and multiplied by a constant ($1x10^{22}$) to bring all values above 1 (see **Methods**). Samples with abundance above the detection limit in both platforms are shown. Each dot denotes one metabolite. The bounds of the

boxes represent the 25th and 75th percentiles, the horizontal center line indicates the median, and the whiskers extend to data points within a maximum of 1.5 times the IQR. Two-sided Wilcoxon signed-rank test was used in **a** and **b**. qPCR data in **b** were log10 transformed before statistical testing.