# Assessment of infant outgrowth of cow's milk allergy in relation to the faecal microbiome and metaproteome

**Diana M. Hendrickx[1,*], Ran An[1,+], Sjef Boeren[2], Sumanth K Mutte[2], PRESTO study team[#], Jolanda M Lambert[3], and Clara Belzer[1]**

[1]Laboratory of Microbiology, Wageningen University, Wageningen, The Netherlands
[2]Laboratory of Biochemistry, Wageningen University, Wageningen, The Netherlands
[3]Danone Nutricia Research, Utrecht, The Netherlands
* diana.hendrickx@wur.nl
[+] current address: Department of Food Science and Technology, School of Agriculture and Biology, Shanghai Jiao Tong University, Shanghai 200240, China.
# A list of authors and their affiliations appears at the end of the paper.

**ABSTRACT**

**Previous studies provide evidence for an association between modifications of the gut microbiota in early life and the development of food allergies.**
**We studied the faecal microbiota composition (16S rRNA gene amplicon sequencing) and faecal microbiome functionality (metaproteomics) in a cohort of 40 infants diagnosed with cow's milk allergy (CMA) when entering the study. Some of the infants showed outgrowth of CMA after 12 months, while others did not. Faecal microbiota composition of infants was analysed directly after CMA diagnosis (baseline) as well as 6 and 12 months after entering the study.**
**The aim was to gain insight on gut microbiome parameters in relation to outgrowth of CMA.**
**The results of this study show that microbiome differences related to outgrowth of CMA can be mainly identified at the taxonomic level of the 16S rRNA gene, to a lesser extent at the taxonomic microbial protein level, and not at the microbial functional protein level. At the 16S rRNA gene level outgrowth of CMA is characterized by lower relative abundance of *Lachnospiraceae* at baseline and lower *Bacteroidaceae* at visit 12 months.**

## Introduction

The first three years of life are a key period for the development of the gut microbiome [1]. In this period, two transition stages can be distinguished [2]. During the first stage, immediately after birth, the gut microbiome is dominated by Proteobacteria and Actinobacteria [3]. During the second stage, when infants start to eat solid foods until about three years of age, an

adult-like microbiome, with Bacteroidetes and Firmicutes as most abundant phyla, is established [4].

Gut microbiome development during the first year of life is strongly influenced by various factors such as mode of delivery (vaginal/Caesarean) [5], antibiotics use [6], mode of feeding (breast fed/formula fed) [7] and lack of siblings [8].

Several studies provide evidence for an association between gut dysbiosis, induced by the factors mentioned above, and the development of food allergies (FA) [9-11]. Previous research has shown that children with allergy have lower faecal DNA levels of *Bacteroides* [12], *Lactobacillus* [12] and *Bifidobacterium* [13] compared to healthy children. Furthermore, increased faecal DNA levels of *Clostridium difficile* [14], *Staphylococcus* [12] and *Escherichia coli* [14] have been associated with an increased risk of developing allergies. In children with cow's milk allergy (CMA), lower faecal DNA levels of Bacteroidetes and higher levels of Proteobacteria were observed [15].

Metaproteomics approaches on faecal samples can provide insights into the functional roles of the intestinal microbiome in human diseases [16]. However, they have seldom be applied to allergic diseases. One study applied metaproteomics on the infant gut microbiome to unravel functional signatures in infants with atopic dermatitis [17]. To the best of our knowledge, metaproteomics has not been used to study the microbiome and its function in CMA.

In this study, we considered both faecal metaproteomics and 16S rRNA gene amplicon sequencing data. Separate analysis was performed on each data type, but selection of core microbiota was based on combined rules applied to the two types of data. We study the changes in the microbiome of children diagnosed with CMA some of whom developed tolerance to cow's milk proteins after one year and others did not. In this way, we aim to identify potential microbial factors associated with developing tolerance to cow's milk.

# Methods

## Sample collection and experimental design

This study included 120 samples of the prospective interventional multicenter study PRESTO (NTR3725), retrieved from Danone Nutricia Research, where infants diagnosed with immunoglobulin E (IgE) mediated CMA received specialized nutrition aimed to induce tolerance to cow's milk. Some infants received an amino acid-based formula (AAF), while others received an AAF with a synbiotic blend (AAF-syn) (oligosaccharides (oligofructose, inulin) + *Bifidobacterium breve M-16V*). Detailed information about the dosage has been described by Chatchatee et al [18]. After 12 months, cow's milk tolerance was not different between infants that were fed with the AAF and those that were fed with the AAF-syn, and was in line with natural outgrowth of CMA [18].

The PRESTO study included 169 subjects, of which 40 were selected for our study based on sample, 16S rRNA gene sequencing and immunological data availability for a study duration up to 12 months. For those 40 infants, samples were collected at three visits, resulting in 120 samples: the baseline visit (0 months) where the subjects entered the study after CMA

diagnosis, 6 months after entering the study and 12 months after entering the study. Faecal microbiota characterization was performed at the three visits for these 40 infants. All infants had been breastfed and were of age 3 - 13 months of age when entering the study (baseline visit). Of these 40 infants, 24 (10 infants in AAF group and 14 infants in AAF-syn group) developed tolerance to cow's milk after 12 months, while 15 infants (6 infants in AAF group and 9 infants in AAF-syn group) were still allergic to cow's milk. For one infant the tolerance to cow's milk status at visit 12 months was inconclusive, and this subject was removed from the analysis. The clinical characteristics of all infants included in this study are shown in Supplementary Table S1. For both metaproteomics and 16S rRNA gene amplicon sequencing 120 samples were collected, resulting in 117 samples after removing those of the infant whose allergy status at 12 months was inconclusive.

## Ethics approval

This multicenter study was designed and conducted in accordance World Medical Association (WMA) Declaration of Helsinki and the International Conference on Harmonization guidelines for Good Clinical Practice [18]. The following national ethics committees, institutional review boards and regulatory authorities approved the study protocol and amendments:

- United Kingdom: NRES Committee North East - Sunderland (Central Ethics Committee MREC) and the local R&Ds from the following hospitals: Great Northern Children's Hospital / Newcastle General Hospital; Southampton General Hospital; Guys & St Thomas; Barts / Royal London Hospital; Leicester Royal Infirmary
- Germany: Ethikkommission Charité – Ethikausschuss 2 am Campus Virchow-Klinikum; Ethikkommission Ärztekammer Nordrhein Düsseldorf, Ethik-Kommission der Medizinischen Fakultät der Ruhr Universität Bochum
- Italy: Comitato Etico per la Sperimentazione Clinica della Province della Pronvincia di Padova; Comitato Etico per la Sperimentazione Clinica della Province di Verona e Rovigo
- Singapore: Singhealth Centralised Institutional Review Board (CIRB) E; National Healthcare Group (NHG) Domain Specific Review Board
- Thailand: Institutional Review Board of the Faculty of Medicine, Chulalongkorn University; Committee on Human Rights Related to Research Involving Human Subjects, Faculty of Medicine Ramathibodi Hospital, Mahidol University; Ethics Committee of the Faculty of Medicine, Prince of Songkla University
- United States of America: Institutional Review Board of the Mount Sinai School of Medicine; Institutional Review Board for Human Subject Research for Baylor College of Medicine and Affiliated Hospitals (BCM IRB); University of Arkansas for Medical Sciences (UAMS) Institutional Review Board

Written informed consent for the collection and analysis of the data was obtained from the parents of all infants included in this study [18].

## 16S-rRNA gene amplicon sequencing

PCR amplification and sequencing of the V3-V4 region of the 16S rRNA gene was performed by LifeSequencing S.L. (Valencia, Spain). The 16S rRNA gene amplicons were sequenced using the 2 x 300 bp paired-end MiSeq protocol (Illumina).

## Sample preparation for metaproteomics by nLC-MS/MS

Total protein was isolated from faeces and subjected to metaproteomics preparation based on protein aggregation capture procedure for mass spectrometry (MS)[19]. Specifically, 10-30 mg faeces was mixed with 125 µl 100 mM Tris pH 8 (Duchefa Biochemie) and sonicated for 15 seconds with an MSE Soniprep 150 Needle sonicator (Beun de Ronde B.V.), at an amplitude of 22 µm. Twenty-five microliter of the slurry was centrifuged and the supernatant was subjected to protein concentration determination using Pierce TM BCA protein assay Kit (Thermo Fisher Scientific). The protein slurry was diluted to 1 ng/µl with 100 mM Tris pH 8. To 60 µl of the 1 ng/µl protein slurry, 150 mM dithiothreitol (Sigma Life Science) was added and incubated at 37 °C gently shaking for 45 min, followed by mixing with 198 µl 8 M urea (Sigma-Aldrich) and 27 µl 200 mM acrylamide (Sigma Life Science), and 30 min room temperature incubation. Subsequently, 4 µl of 10% trifluoroacetic acid (TFA, Alfa Aesar Chemicals), 8 µl of SpeedBeads™ magnetic carboxylate modified particles (50% GE Healthcare 45152105050250 and 50% Thermo Scienfific 65152105050250 after two times washing with milli-Q water) and 750 µl acetonitrile (Biosolve B.V.) were added to the mixture, followed by 20 min room temperature incubation. The liquid was removed and beads with protein were washed twice with 1 ml 70% ethanol and with 1 ml 100% acetonitrile. Afterwards, beads with protein were subjected to overnight digestion by adding 100 µl 5 ng/µl sequencing grade trypsin (Roche Diagnostic GmbH) in ammonium bicarbonate (Sigma-Aldrich). Digestion was performed at room temperature while gently shaking overnight. The digestion was stopped by adding 4 µl 10% TFA to the mixture. The liquid was removed from the beads and transferred to a C18 µcolumn (description below). The remaining pellet was washed with 100 µl 1ml/L formic acid (Biosolve B.V.), which was also transferred to the µcolumn. After eluting all liquid through the µcolumn (while keeping the membrane not dry), 15 µl (1:1) formic acid & acetonitrile solution was added to the µcolumn and eluted the liquids. The filtered solution was then concentrated (Eppendorf Concentrator Plus) to a final volume of 10-15 µl, which volume was adjusted to exactly 50 µl with 1ml/L formic acid before storage in the -20°C freezer until analysis. The C18 µcolumn was made by adding two 1 mm pieces of C18 disk (Affinisep AttractSPE™ Disk Bio C18), 200 µl methanol (Hipersolv Chromanorm) and 4 µl 50% µcolumn material (Lichroprep RP-18) in methanol into a 200 µl pipette tip, which was eluted and washed with 100 µl methanol and equilibrated with 100 µl 1ml/L formic acid before usage.

## Metaproteomics nLC-MS/MS

Five microliter of peptide sample (defrozen and centrifuged at 12.000 * g for 30 min) was loaded directly onto a 0.10 * 250 mm ReproSil-Pur 120 C18-AQ 1.9 µm beads analytical column (prepared in-house) at a constant pressure of 825 bar (flow rate of circa 650 nL/min) with 1 ml/l HCOOH in water and eluted at a flow of 0.5 µl/min with a 50 min linear gradient from 9% to 34% acetonitril in water with 1 ml/l formic acid with a Thermo EASY nanoLC1000.
An electrospray potential of 3.5 kV was applied directly to the eluent via a stainless steel needle fitted into the waste line of a micro cross that was connected between the nLC and the analytical column.
A Field Asymmetric Ion Mobility Spectrometry (FAIMS) setup was used at a set compensation voltage of -45 V to increase the number of MSMS spectra obtained from doubly charged peptides. Full scan positive mode Fourier transform mass spectrometry (FTMS) spectra were measured between m/z 380 and 1400 on a Exploris 480 (Thermo electron, San Jose, CA, USA) at resolution (60000). MS and MSMS AGC targets were set to 300%, 100% respectively or maximum ion injection times of 50 ms (MS) and 30 ms (MSMS) were used. Higher-energy Collision dissociation (HCD) fragmented (Isolation width 1.2 m/z, 28% normalized collision energy) MSMS scans of the 25 most abundant 2-5+ charged peaks in the MS scan were recorded in data dependent mode (Resolution 15000, threshold 2e4, 15 s exclusion duration for the selected m/z +/- 10 ppm).

## 16S rRNA amplicon sequencing analysis

After demultiplexing the read pairs, low-quality sequences were removed by trimming the reads using a quality score (Q-score) threshold of 20. The trimmed reads were merged using PEAR [20], and merged reads with a minimal length of 300 were retained if they had a Q-score larger than 25 over a window of 15 bases and no ambiguous bases were present. Dereplication and counting of the merged reads was performed using mothur [21], and reads with < 2 reads over all samples were removed. Next, VSEARCH [22] together with the RDP gold database [23] were used to eliminate chimeras. Filtering of reads including PhiX and Adapter sequences (as defined in Deblur [24]) was performed using QIIME2 [25]. For each amplicon sequence variant (ASV), taxonomy was assigned at the genus level using the Ribosomal Database Project (RDP) [26] classifier against the SILVA 138 reference database [27], resulting in a classification of the ASVs into 173 genera.

## Database construction

ASVs were identified at the species level with NCBI nucleotide BLAST [28], using the 16S ribosomal RNA sequences (Bacteria and Archaea) database. The maximum number of aligned sequences was set to 10. For all other parameters the default settings were used.

For the resulting 327 species the non-redundant proteomes were obtained from UniProtKB [29]. In case a species had no proteomes in UniProtKB, the NCBI proteome was obtained. For taxa with more than 2 proteomes in UniProtKB, two representative proteomes were selected. In case there was a reference proteome, this proteome was included and the second proteome was selected based on the Complete Proteome Detector (CPD) [29] and Benchmarking Universal Single-Copy Ortholog (BUSCO) [30] data completeness and quality scores. The resulting list of proteomes is provided in Supplementary Table S2. Including too many rare taxa found by 16S rRNA sequencing in a proteomics database can lead to a lower number of protein identifications. Therefore, a threshold for removing spurious taxa had to be defined. Genera were ordered from high to low relative abundance and multiple databases were constructed including the top genera that covered 50%, 66%, 75%, 80%, 90%, 92% and 100% of the total relative abundance in the 16S rRNA sequencing data. The database constructed using a threshold of 92% for the total relative abundance resulted in the highest number of protein identifications. To keep the database size limited yet complete, only the identified proteins (instead of complete proteomes) from the remaining 8% of the species were added to the database generated in the previous step. Since many proteins can be identical/redundant across multiple species/proteomes, we reduced the redundancy by clustering proteins that were identical using CD-HIT [31] with settings '-c 1.0 -d 500 -M 0'. The resulting non-redundant microbial proteomes database contained 416437 proteins.

## Metaproteomics data analysis

The obtained nLC-MS/MS spectra were analysed with MaxQuant version 1.6.3.4 [32] using the "Specific Trypsin/P" Digestion mode with maximally 2 missed cleavages and further default settings for the Andromeda search engine [33] (first search 20 ppm peptide tolerance, main search 4.5 ppm tolerance, MSMS fragment match tolerance of 20 ppm). Propionamide (C) was set as a fixed modification, while variable modifications were set for Protein N-terminal Acetylation and M oxidation with the maximum number of modifications per peptide set to 5. Next to the metaproteomics database described above under "database construction", a Human database downloaded from Uniprot (UP000005640 with 20400 sequences) and a contaminants database with 66 sequences of common contaminants were used. The "label-free quantification" as well as the "match between runs" options were enabled. De-amidated peptides were allowed to be used for protein quantification and all other quantification settings were kept default.
Proteins identified by only one or no unique peptide, reverse hits and hits only identified by site were filtered out. In this way we obtained 1820 protein groups, of which 1653 were from bacteria. Before extra filtering, the protein groups included 9085 proteins in total, of which 8797 were bacterial proteins.

## Statistical analysis of 16S rRNA gene sequencing and metaproteomics data

All analyses (except LEfSe, see below) were performed in R version 3.6.1 [34]. Clinical factors associated with outgrowth of cow's milk allergy were determined by a two-sided Mann-Whitney U-test for numeric variables and a Fisher's exact test for binary variables.
For the microbiota analysis, ASVs were converted to relative abundances at the family level using the R microbiome package [35] version 1.8.0. For the metaproteomics data, relative abundances were calculated from the intensity Based Absolute Quantifcation (iBAQ) intensities.
Core microbiota were determined at the family level using the R microbiome package. Core taxa were defined as taxa that have a relative abundance higher than 1% in at least 50% of the 16S rRNA sequencing samples or at least 50% of the metaproteomics samples.
Bacterial proteins were functionally annotated by assigning protein identifiers to Kyoto Encyclopedia of Genes and Genomes (KEGG) [36] orthology (KO) identifiers, and classifying the KO identifiers using KEGG Brite hierarchy level C.
To compare relative abundances of core taxa, top 10 microbial protein functional classes and top 10 human protein classes within allergy groups and between visits, the data were normalized by centred log ratio (CLR) transform and a linear mixed model (LMM) analysis, with outgrowth of CMA, visit and outgrowth of CMA x visit as fixed effects and subject as random effect, was performed for each core taxon or protein class. The interaction term outgrowth of CMA x visit was included as we expected that visits compare differently at the level of outgrowth of CMA. The LMM was also used to determine differences in core taxa between allergy groups within each visit. P-values for pairwise comparisons were calculated from the model, and adjusted for multiple testing using the Benjamini-Hochberg procedure [37].
To relate variability in the microbiome to environmental and clinical factors, redundancy analysis (RDA) was performed. Aitchison distance [38] (Euclidian distance on CLR transformed data), which has been proposed as a suitable distance metric for compositional data [39], was used for RDA. RDA analysis was performed with the R vegan package [40] version 2.5.-6, using stepwise forward and backward model selection where a variable is added when its p-value is smaller than or equal to 0.05, and removed when its p-value is larger than 0.1. Furthermore, partial RDA with outgrowth of CMA as explanatory variable and the remainder of the environmental variables as covariates was performed to determine the proportion of variance in the microbiome explained by outgrowth of CMA. A similar RDA and partial RDA analysis was performed to relate variability in the microbial proteome and human proteins to environmental factors. Furthermore, another RDA was conducted to relate variability in the proteome and metaproteomics-based microbial composition to human proteins.
Spearman correlation between 16S gene- and protein-based taxonomic classification, and between human and microbial proteins, were calculated. P-values were determined with Monte Carlo permutation (10 000 permutations), and correlations with a p-value below 0.05 were considered significant.

Discriminative microbial features between allergy groups (outgrowth versus no outgrowth of CMA at 12 months) and between visits were determined by Linear discriminant analysis Effect Size (LEfSe) analysis [41], using the Galaxy module (http://huttenhower.sph.harvard.edu/galaxy). An alpha value of 0.05 for the factorial Kruskal-Wallis test among classes and a threshold of 2.0 on the logarithmic linear discriminant analysis (LDA) score for discriminative features were used for the calculations. To indicate which samples belong to the same subject in the pairwise comparisons between visits, a subject variable was included in the analysis.

# Results

## Environmental and clinical factors associated with outgrowth of CMA

To determine environmental and clinical factors associated with outgrowth of CMA in our data set, we performed statistical tests (two-sided Mann-Whitney U-test for numeric variables, Fisher's exact test for binary variables) to the metadata of the subjects used in this study (clinical characteristics, Supplementary Table S1).
Among infants with persistent CMA, a significantly higher proportion of infants with parental allergy is observed compared to infants that show outgrowth of CMA (Table S3 and Figure 1A). This suggests that persistent CMA is more common in case of parental allergy.
At baseline visit, a significantly lower score for atopic dermatitis (SCORAD) was observed in infants that show outgrowth of CMA compared to infants that do not (Table S3 and Figure 1B). The SCORAD did not change significantly over time in any of the two allergy groups (Figure 1C).

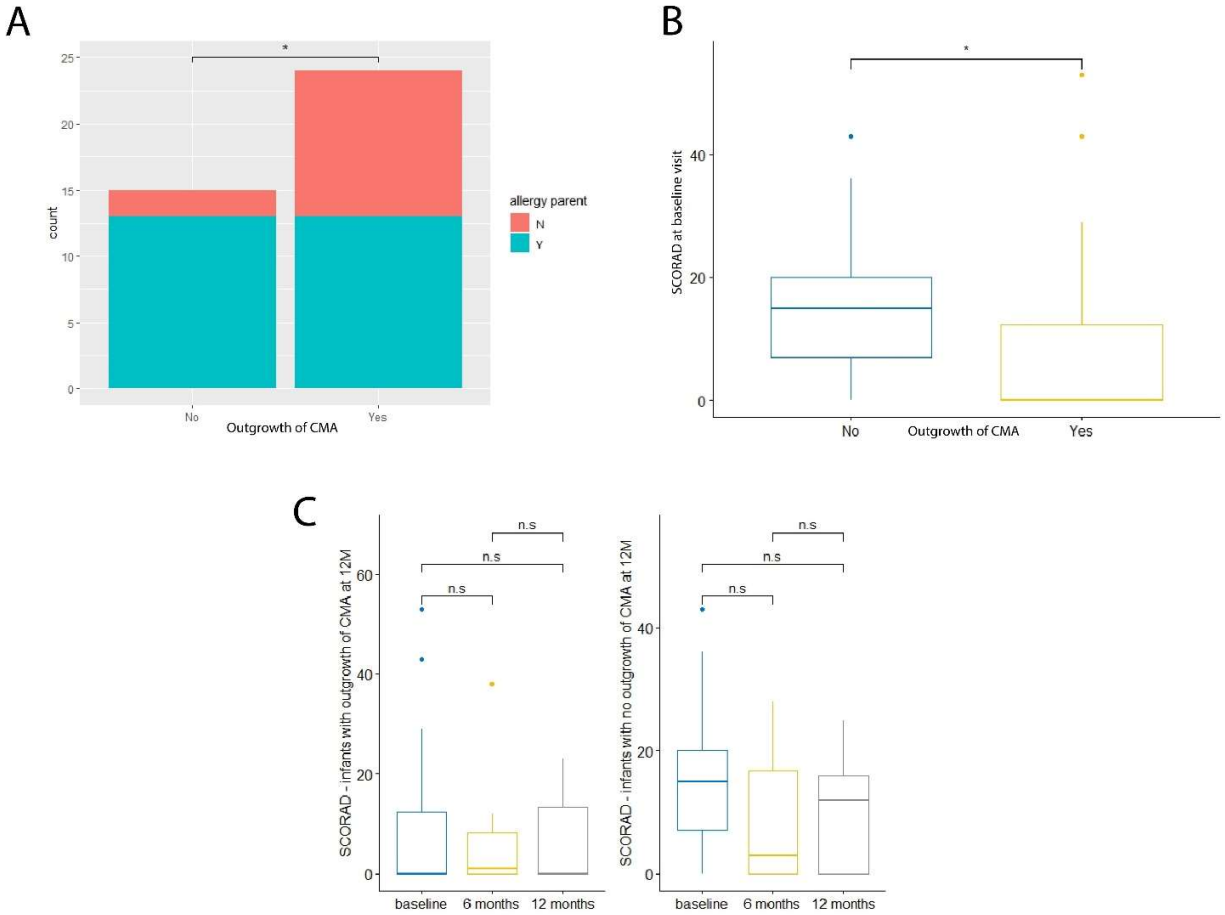


Figure 1. (A) Stacked bar plot for parental allergy. No = no outgrowth of CMA at visit 12 months; Yes = outgrowth of CMA at visit 12 months. (B) Boxplot for SCORAD at baseline visit. No = no outgrowth of CMA at visit 12 months; Yes = outgrowth of CMA at visit 12 months. (C) Boxplots for SCORAD over time. Left: outgrowth of CMA at 12M; right: no outgrowth of CMA at 12M. n.s.: not significant.

## The faecal metaproteome is dominated by bacterial proteins from *Bifidobacteriaceae* and *Lachnospiraceae*

Analysis of metaproteomics data was done by grouping the subjects into CM tolerant or not tolerant at visit 12 months. The different visits were also analysed separately. For all groups, more than 50% of the total proteins are bacterial, the remaining proteins are human proteins (Figure 2A). The microbiota composition at the family level by metaproteomics showed high levels of *Bifidobacteriaceae* and *Lachnospiraceae* proteins (Figure 3A). Among the human proteins immunoglobulins are highly abundant (Figure 2B).

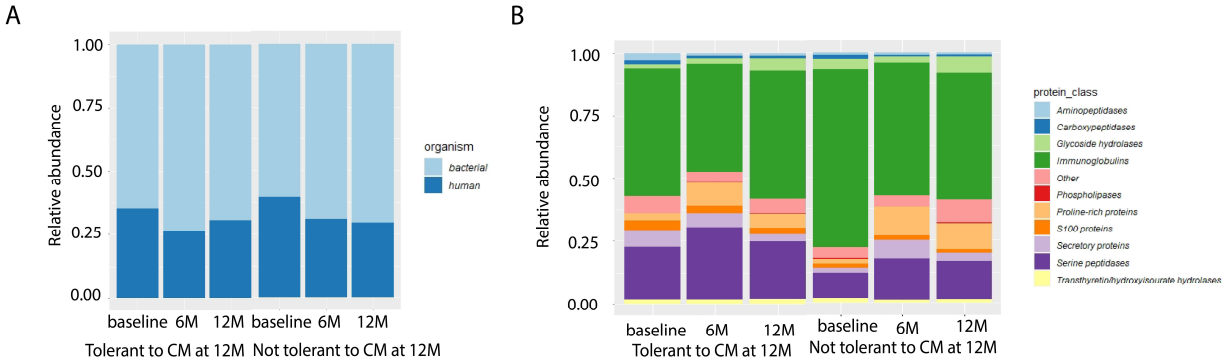


Figure 2. Ratio human/bacterial proteins (A) and human proteins composition (B) at each visit for the group that became tolerant to cow's milk (CM) at visit 12 months (12M) and the group that was still allergic to CM at 12M.

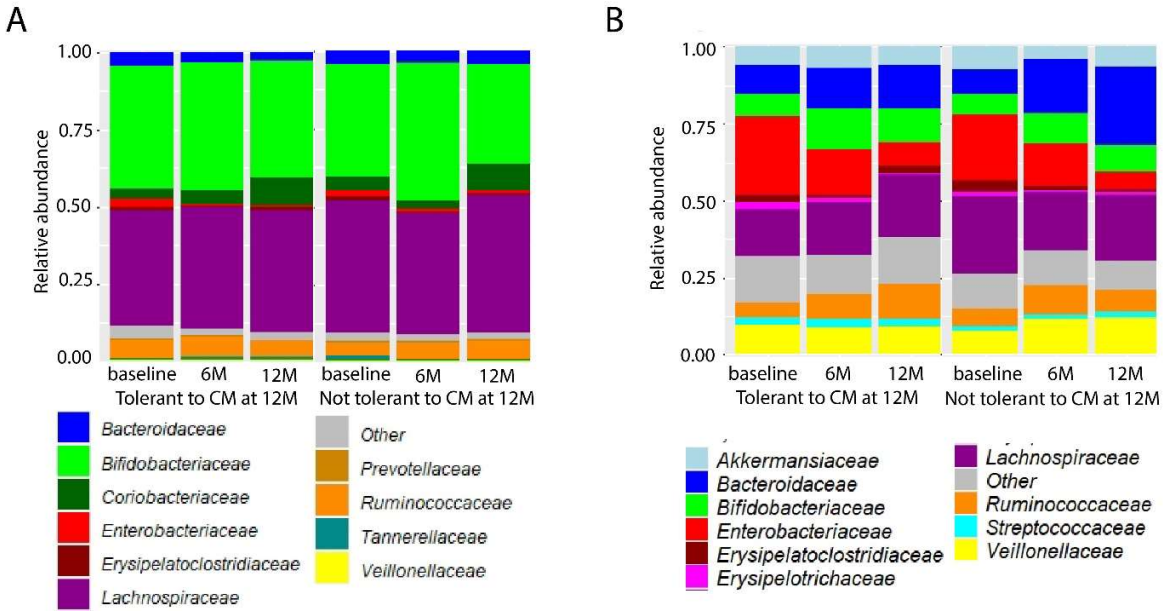


Figure 3. Metaproteomics-based taxonomic profiles (family level) (A) and 16S rRNA gene-based taxonomic profiles (B) at each visit for the group that became tolerant to cow's milk (CM) at visit 12 months (12M) and the group that was still allergic to CM at 12M.

## Differences and similarities between 16S rRNA gene-based and metaproteomics-based taxonomic profiles

Metaproteomics- and 16S rRNA gene-based taxonomic profiles of the top 10 most abundant families across all samples were determined and compared between allergy groups (CM tolerant versus non-tolerant at visit 12 months) and visits (Figures 3A and B).

329 ***Top 10 most abundant taxa in 16S rRNA gene and metaproteomics data***

330 The top 10 most abundant families have 7 taxa in common between 16S rRNA gene-based
331 and metaproteomics-based taxonomic profiles. They are: *Bacteroidaceae*,
332 *Bifidobacteriaceae*, *Enterobacteriaceae*, *Erysipelatoclostridiaceae*, *Lachnospiraceae*,
333 *Ruminococcaceae* and *Veillonellaceae* (Figure 3A and B). *Coriobacteriaceae*, *Prevotellaceae*
334 and *Tannerellaceae* were only among the top 10 taxa of metaproteomics-based taxonomic
335 profiles (Figure 3A). Taxa that were only among the top 10 in 16S rRNA gene-based profiles
336 were *Akkermansiaceae*, *Erysipelotrichaceae* and *Streptococcaceae* (Figure 3B). Both 16S
337 rRNA-based and metaproteomics based taxonomic profiles show high inter-individual
338 variation (Figures S1 and S2).

339

340 ***Correlation core taxa between 16S rRNA gene-based and metaproteomics-based***
341 ***profiles***

342 Core taxa, defined as taxa that have a relative abundance higher than 1% in at least 50% of
343 the 16S rRNA gene sequencing or metaproteomics samples included *Bifidobacteriaceae*,
344 *Coriobacteriaceae*, *Bacteroidaceae*, *Lachnospiraceae*, *Ruminococcaceae*, *Veillonellaceae* and
345 *Enterobacteriaceae*. *Bifidobacteriaceae*, *Bacteroidaceae*, *Lachnospiraceae* and
346 *Ruminococcaceae* fulfil these conditions for both 16S rRNA gene sequencing and
347 metaproteomics samples, *Veillonellaceae* and *Enterobacteriaceae* only for 16S rRNA gene
348 sequencing and *Coriobacteriaceae* only for metaproteomics. The relative abundance of
349 *Bifidobacteriaceae*, *Bacteroidaceae*, *Lachnospiraceae*, *Ruminococcaceae*, *Veillonellaceae* and
350 *Enterobacteriaceae* correlated significantly (p-value ≤ 0.05) between 16S rRNA-based and
351 metaproteomics-based profiles (Table 1). However, apart from *Bifidobacteriaceae*, the
352 correlations were low.

353 Table 1. Spearman correlation between 16S rRNA gene- and protein-based taxonomic
354 classification. P-values were determined with Monte Carlo permutation (10 000
355 permutations). P-values below 0.05 are considered significant.

| **Family** | **Spearman correlation** | **p-value** |
|---|---|---|
| *Bifidobacteriaceae* | 0.840 | **< 1e-04** |
| *Coriobacteriaceae* | 0.139 | 0.1316 |
| *Bacteroidaceae* | 0.250 | **0.0059** |
| *Lachnospiraceae* | 0.563 | **< 1e-04** |
| *Ruminococcaceae* | 0.348 | **2e-04** |
| *Veillonellaceae* | 0.190 | **0.0388** |
| *Enterobacteriaceae* | 0.313 | **4e-04** |

356

357

358

## Significant differences between allergy groups (outgrowth versus no outgrowth of CMA) in 16S rRNA gene-based relative abundance levels

To determine microbial variables most likely explaining differences between allergy groups in the 16S rRNA gene sequencing data, Linear Mixed Model (LMM) analysis on the core taxa and LEfSe analysis on the whole data set were performed. Below, differences at the family level are reported in detail, for differences at the higher taxonomy levels we refer to Figure 4.

### ***Significant 16S rRNA gene-based differences between allergy groups within each visit***

The results of LMM analysis showed significantly higher relative abundances of *Lachnospiraceae* at baseline in infants with persistent CMA compared to infants who outgrew their CMA (Table S4). Furthermore, according the LEfSe analysis the families *Carnobacteriaceae*, *Lachnospiraceae* and the *Peptostreptococcales-Tissierellales* family have significantly higher relative abundances at baseline visit in the group with persistent CMA (Figure 4A). In the group with outgrowth of CMA, higher relative abundances of *Enterococcaceae* and *Leuconostocaceae* were observed (Figure 4A). At visit 6 months, higher relative abundances of *Aerococcaceae* and *Pasteurellaceae* were observed in the group with persistent CMA, while in the group with outgrowth of CMA, higher relative abundances of the family *Eubacteriaceae* were observed (Figure 4B). At visit 12 months, higher relative abundances of *Bacteroidaceae* were observed in the infants with persistent CMA , while the infants who outgrew their CMA showed higher relative abundances of the families *Rikenellaceae*, *Actinomycetaceae* and *Eggerthellaceae* (Figure 4C).

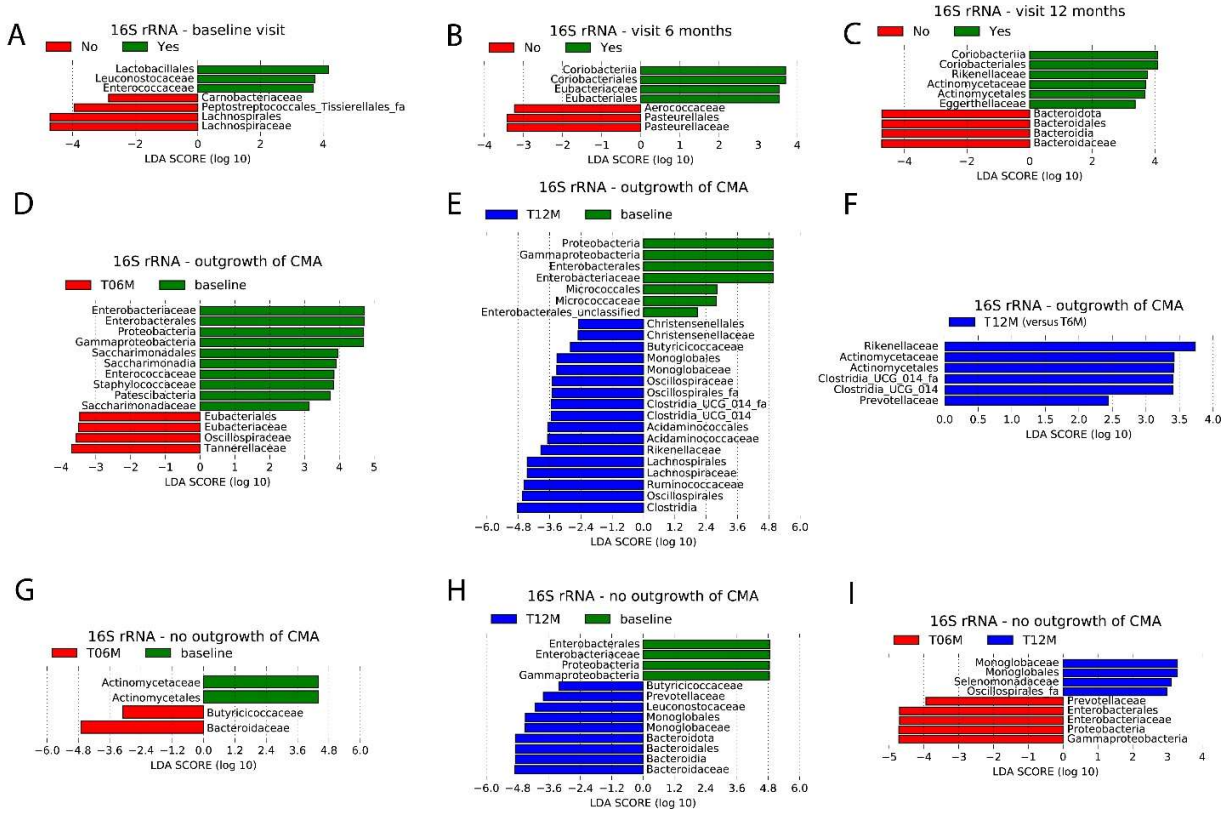


Figure 4. LEfSe analysis of 16S rRNA-based relative abundances at the family level using an alpha value of 0.05 for the factorial Kruskal-Wallis test among classes and a threshold of 2.0 on the logarithmic LDA score for discriminative features. (A-C) Using tolerance to cow's milk (CM) at visit 12 months as class, Yes: tolerant to CM at 12 months; no: not tolerant to CM at 12 months. Plots of discriminative features at baseline visit (A), visit 6 months (B) and visit 12 months (C). (D-F) Pairwise comparison between visits within the group with outgrowth of CMA, using visit as class. (D) Visit 6 months (T06M) versus baseline; (E) visit 12 months (T12M) versus baseline; (F) visit 12 months (T12M) versus visit 6 months (T06M). (G-I) Pairwise comparison between visits within the group with persistent CMA, using visit as class. (G) Visit 6 months (T06M) versus baseline; (H) visit 12 months (T12M) versus baseline; (I) visit 12 months (T12M) versus visit 6 months (T06M).

***Significant 16S rRNA gene-based differences between visits within allergy groups***

When comparing differences between visits within each allergy group, LMM analysis showed significantly higher relative abundances of *Ruminococcaceae* at visit 12 months compared to baseline visit in the group with outgrowth of CMA (Table S5). In the group with persistent CMA, relative abundances of *Bacteroidaceae* were significantly higher at visit 12 months compared to baseline visit (Table S5). Significant differences of *Enterobacteriaceae* were observed in both allergy groups. For the group with outgrowth of CMA, significantly lower relative abundances of *Enterobacteriaceae* were observed at visit 12 months compared to

404 baseline visit, as well as at visit 6 months compared to baseline visit (Table S5). In the
405 group with persistent CMA, *Enterobacteriaceae* were only lower at visit 12 months compared
406 to baseline (Table S5).
407 In the group with outgrowth of CMA, the LEfSe analysis showed higher relative abundances
408 of the families *Eubacteriaceae*, *Oscillospiraceae* and *Tannerellaceae*, and lower relative
409 abundances of *Enterobacteriaceae*, *Enterococcaceae*, *Staphylococcaceae* and
410 *Saccharimonadaceae* at visit 6 months compared to baseline (Figure 4D). Higher relative
411 abundances of the families *Christensenellaceae*, *Butyricicoccaceae*, *Monoglobaceae*,
412 *Oscillospiraceae*, *Oscillospirales* family, *Clostridia UCG 014* family, *Acidaminococcaceae*,
413 *Rikenellaceae*, *Lachnospiraceae* and *Ruminococcaceae*, and lower relative abundances of
414 *Enterobacteriaceae* and *Micrococcaceae* were observed at visit 12 months compared to
415 baseline visit (Figure 4E). Higher relative abundances of the families *Rikenellaceae*,
416 *Actinomycetaceae* and *Prevotellaceae* were observed at visit 12 months compared to visit 6
417 months (Figure 4F).
418 At the family level, LEfSe analysis for the group with persistent CMA showed significantly
419 higher relative abundances of *Butyricicoccaceae* and *Bacteroidaceae*, and significantly lower
420 relative abundances of *Actinomycetaceae* at visit 6 months compared to baseline (Figure
421 4G). Higher relative abundances of the families *Butyricicoccaceae*, *Prevotellaceae*,
422 *Leuconostocaceae*, *Monoglobaceae* and *Bacteroidaceae*, as well as lower relative abundances
423 of *Enterobacteriaceae* were observed at visit 12 months compared to baseline (Figure 4H).
424 Moreover, higher relative abundance levels of the families *Monoglobaceae* and
425 *Selenomonadaceae*, as well as lower relative abundance levels of *Prevotellaceae* and
426 *Enterobacteriaceae* were observed at visit 12 months compared to visit 6 months (Figure
427 4I).
428

## 429 Significant differences between allergy groups (outgrowth versus no 430 outgrowth of CMA) in metaproteomics-based relative abundance 431 levels

432 Linear Mixed Model (LMM) and LEfSe analysis were performed to determine microbial
433 variables most likely explaining differences between allergy groups in the metaproteomics
434 data. Below, differences at the family level are reported in detail, for differences at the
435 higher taxonomy levels we refer to Figure 5.

436

### 437 *Significant metaproteomics-based differences between visits within allergy groups*

438 The LMM analysis did not identify any significant differences in core taxa between allergy
439 groups within visits (Table S6). The results of the LEfSe analysis at baseline show higher
440 relative abundances of *Eggerthellaceae* proteins in infants with persistent CMA compared to
441 infants who outgrew their CMA (Figure 5A). For the other visits, no significant differences
442 between allergy groups were observed.

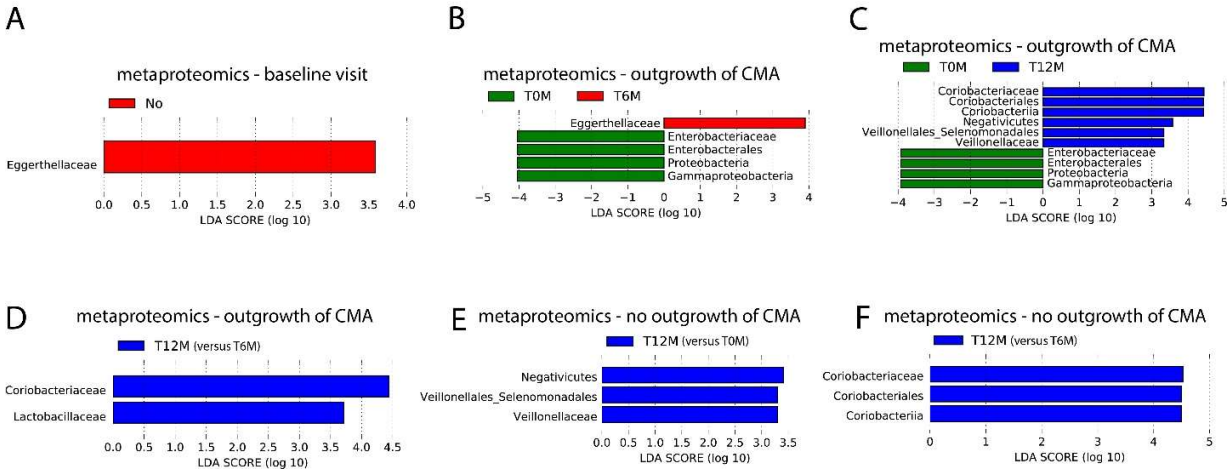


Figure 5. LEfSe analysis of metaproteomics-based relative abundances at the family level using an alpha value of 0.05 for the factorial Kruskal-Wallis test among classes and a threshold of 2.0 on the logarithmic LDA score for discriminative features. (A) Using tolerance to cow's milk (CM) at visit 12 months as class, Yes: tolerant to CM at 12 months; no: not tolerant to CM at 12 months. Plot of discriminative features at baseline visit. (B-D) Pairwise comparison between visits within the group with outgrowth of CMA, using visit as class. (B) Visit 6 months (T6M) versus baseline (T0M); (C) visit 12 months (T12M) versus baseline (T0M); (D) visit 12 months (T12M) versus visit 6 months (T6M). (E-F) Pairwise comparison between visits within the group with persistent CMA, using visit as class. (E) visit 12 months (T12M) versus baseline (T0M); (F) visit 12 months (T12M) versus visit 6 months (T6M).

***Significant metaproteomics-based differences between visits within allergy groups***

In infants who outgrew their CMA, the LMM analysis comparing visits identified significantly lower relative abundance of *Enterobacteriaceae* proteins at visit 6 months compared to the baseline visit (Table S7). No significant differences between visits were obtained by the LMM analysis in the group with persistent CMA.

In the group with outgrowth of CMA, the LEfSe analysis showed significantly higher relative abundances of *Eggerthellaceae* proteins, and significantly lower relative abundances of *Enterobacteriaceae* proteins at visit 6 months compared to baseline visit (Figure 5B). Furthermore, significantly higher relative abundances of *Coriobacteriaceae* and *Veillonellaceae* proteins, as well as significantly lower relative abundances of *Enterobacteriaceae* proteins were observed at visit 12 months compared to baseline visit (Figure 5C). Moreover, significantly higher relative abundances of *Coriobacteriaceae* and *Lactobacillaceae* proteins were identified at visit 12 months compared to visit 6 months (Figure 5D).

The LEfSe analysis comparing visits within the group with persistent CMA did not identify any significant differences between baseline visit and visit 6 months. Significantly higher relative abundances of *Veillonellaceae* proteins were observed at visit 12 months compared

472 to the baseline visit (Figure 5E). Moreover, significantly higher relative abundances of
473 *Coriobacteriaceae* proteins were observed at visit 12 months compared to visit 6 months
474 (Figure 5F).

475

## 476 Significant differences between allergy groups (outgrowth versus no 477 outgrowth of CMA) in relative abundance levels of microbial protein 478 functional classes

479 The results of Linear Mixed Model (LMM) analysis, applied to the top 10 microbial protein
480 functional classes, showed no significant difference between allergy groups within visits
481 (Table S8). Moreover, no significant differences between visits were found within allergy
482 groups (Table S9).

483

## 484 Significant differences between allergy groups (outgrowth versus no 485 outgrowth of CMA) in relative abundance levels of human protein 486 classes

487 To determine if human protein classes could explain differences between allergy groups,
488 Linear Mixed Model (LMM) analysis was applied. The results show no significant differences
489 between allergy groups within visits (Table S10). Furthermore, relative abundances of
490 proline-rich proteins (total of 3 proteins, belonging to 3 protein groups) are significantly
491 higher at visit 12 months compared to the baseline visit in both allergy groups (Table S11).
492 Moreover, relative abundances of immunoglobulins (total of 55 proteins, belonging to 26
493 protein groups) are significantly lower at visit 12 months than baseline in infants with
494 persistent CMA, but not in infants who outgrew their CMA.

495

## 496 Correlation between human and microbial proteins

497 To determine correlation between human and microbial proteins, Spearman correlations
498 between human proteins of the top 10 protein classes and microbial proteins of the core
499 taxa were calculated. Several significant correlations, of which the majority are weak, were
500 found (Table S12). Significant moderate correlations (> 0.4) were observed between
501 proline-rich proteins and *Veillonellaceae* (correlation = 0.4159, p-value < 1e-04),
502 carboxypeptidases and *Veillonellaceae* (correlation =0.4997, p-value < 1e-04), and
503 Transthyretin / hydroxyisourate hydrolases and *Coriobacteriaceae* (correlation =0.5205, p-
504 value < 1e-04).

505

506 **Outgrowth of CMA has a low but significant contribution to the**
507 **variation in 16S-rRNA gene based microbial signatures at visit 6**
508 **months**

509 To determine the contribution of environmental and clinical factors to the variation in the gut
510 microbiota, microbial proteome and human proteins, redundancy analysis (RDA) was
511 performed. Outgrowth of CMA significantly explains variation in 16S rRNA gene based
512 taxonomic profiles at visit 6 months (Table S13). However, compared to other significant
513 features, the contribution of outgrowth of CMA is low (shorter edge in RDA plot, see Figure
514 S3). Therefore, ordination plots coloured by outgrowth of CMA do not show a clear
515 separation between the subjects with outgrowth of CMA and those with persistent CMA.
516 To determine the contribution of outgrowth of CMA to the variation in the gut microbiota and
517 microbial proteome after adjusting for the other environmental variables, partial RDA was
518 performed. The results show that the contribution is low and only significant for the 16S
519 rRNA gene-based taxonomy (% variance explained by outgrowth of CMA = 6.37%, adjusted
520 p-value = 0.043). The partial RDA plot shows that after removing the effect of other
521 environmental factors, a more clear separation between the allergy classes (outgrowth vs
522 persistent CMA) could be observed (Figure S4).
523 For the other visits, the microbial proteome profiles, the protein-based taxonomic profiles,
524 the microbial protein functional profiles and the human protein profiles, the contribution of
525 outgrowth of CMA to the variation in the data was not significant (Tables S13 and S14).

526

527

528 **Contribution of human proteins on variation in the microbial**
529 **proteome**

530 To determine human proteins that significantly explain variation in microbial proteome and
531 protein-based taxonomy profiles, another RDA analysis was performed. Table S15 shows
532 that a considerable fraction of the variance in the microbial proteome and protein-based
533 taxonomy profiles can be explained by human proteins at all visits. No clear grouping of the
534 samples by allergy group was observed in any of the RDA plots (Figure S5).

535

536 # Discussion

537 In this study, 16S rRNA gene amplicon sequencing and metaproteomics were applied on
538 faecal samples in a cohort of infants with CMA at baseline visit, of which some showed
539 outgrowth of CMA after 12 months and others did not. The aim was to get insight in
540 outgrowth of CMA.
541 Apart from analysing the microbiome, environmental and clinical factors associated with
542 outgrowth of CMA were determined. Our results show that among infants with persistent

543 CMA, a significantly higher proportion of infants with parental allergy was observed
544 compared to infants that show outgrowth of CMA. This is consistent with previous findings of
545 Xinias et al., who showed that the risk of not acquiring tolerance to cow's milk at 1 year of
546 age significantly increased in infants with a family history of atopy [42]. Furthermore, a
547 significantly lower SCORAD at baseline was observed in infants that show outgrowth of CMA
548 compared to infants that do not. However, the SCORAD of the majority of the infants in this
549 study is lower than 25, corresponding to mild severity of atopic dermatitis [43]. This suggests
550 that in this cohort, CMA does not manifest mainly through atopic symptoms.
551 The faecal metaproteome was dominated by bacterial proteins from *Bifidobacteriaceae* and
552 *Lachnospiraceae*. Proteins for these two families were also observed in a metaproteomics
553 study of Kingkaw et al. [17] on atopic dermatitis, but the fraction of these taxa was lower than
554 in our study. There are two possible explanations for the differences between our study and
555 Kingkaw et al. First, the infants in our study have another CMA phenotype than atopic
556 dermatitis. Second, the protein identification and quantification of Kingkaw et al. was based
557 on a protein database derived from earlier studies of the gut microbiome which included only
558 10 bacterial families. In our study, we constructed an in-house proteomics database based
559 on the 16S rRNA gene amplicon sequencing data consisting of 19 bacterial families. Third,
560 the cohort in the study of Kingkaw et al. included both healthy and allergic children, while
561 our study includes only children with CMA at baseline. The gut microbiome of infants with
562 food allergy in their first year of life is known to be dominated by *Lachnospiraceae*, while this
563 family is less abundant in the healthy gut microbiome [44].
564 When comparing metaproteomics- and 16S rRNA gene-based taxonomic profiles, the top 10
565 most abundant families have 7 taxa in common: *Bacteroidaceae*, *Bifidobacteriaceae*,
566 *Enterobacteriaceae*, *Erysipelatoclostridiaceae*, *Lachnospiraceae*, *Ruminococcaceae* and
567 *Veillonellaceae*. With the exception of *Erysipelatoclostridiaceae*, these are all common
568 constituents of the healthy gut microbiome [45,46]. *Erysipelatoclostridiaceae* are known to be
569 pathogenic, and their presence in this cohort can be related to the fact that the majority of
570 the infants had one or more infections during the study period. *Coriobacteriaceae*,
571 *Prevotellaceae* and *Tannerellaceae* were only among the top 10 taxa of metaproteomics-
572 based taxonomic profiles. The first two are common taxa in the gut microbiome [45]. The
573 presence of increased levels of *Tannerellaceae* in the gut microbiome has been associated
574 with the administration of synbiotics [47], and its appearance in the top 10 taxa in this study
575 can be explained by the inclusion of 23 infants (59% of the infants) which received a
576 formula with synbiotics. Taxa that were only among the top 10 in 16S rRNA gene-based
577 profiles were *Akkermansiaceae*, *Erysipelotrichaceae* and *Streptococcaceae*. The latter two
578 are common constituents of the gut microbiome [45], while increased levels of
579 *Akkermansiaceae* have been related with synbiotics interventions [47]. Differences between
580 16S rRNA gene-based and metaproteomics-based microbial composition have been observed
581 in a previous study [48], for which two possible explanations were given by the authors. First,
582 it could be that the 16S rRNA gene-based microbial composition is not representative for
583 microbial protein abundance. Second, 16S rRNA gene sequences could be highly similar,
584 which could have led to misclassifications of ASVs.

585 Linear mixed model (LMM) analysis of core taxa and LEfSe analysis on the whole data were
586 applied on the 16S rRNA gene sequencing and metaproteomics data to compare infants who
587 developed tolerance to cow's milk with infants with persistent CMA.
588 The results of both LMM and LEfSe analysis on the 16S rRNA gene sequencing data showed
589 significantly higher relative abundances of *Lachnospiraceae* at baseline in infants with
590 persistent CMA compared to infants who outgrew their CMA. In contrast, a previous study
591 reported higher levels of *Lachnospira pectinoschiza* in infants that developed tolerance to
592 allergy [49]. However, this study focuses on children with a mean age of 57.8 months at
593 baseline, while the infants in our study were 15–25 months old at visit 12 months (3-13
594 months old at baseline). Furthermore, higher relative abundances of *Carnobacteriaceae* and
595 the *Peptostreptococcales-Tissierellales* family were reported by LEfSe at baseline in infants
596 with persistent CMA. To the best of our knowledge, these two families have not been related
597 to CMA in earlier studies. In infants who outgrew their CMA, higher relative abundances of
598 *Enterococcaceae and Leuconostocaceae* at baseline visit were observed. Higher levels of the
599 genus *Enterococcus* have been reported in non-allergic infants when compared to allergic
600 infants [50]. At visit 6 months, infants with persistent CMA showed higher relative abundances
601 of *Aerococcaceae* and *Pasteurellaceae*. This is in contradiction with Kourosh et al.[51], who
602 observed lower Pasteurellales in the allergic group compared to the control. However, this
603 study was not restricted to early life but ranged from birth to the age of 18 years old.
604 Infants who outgrew their CMA showed higher relative abundances of *Eubacteriaceae* at visit
605 6 months. Bunyavanich et al.[52] reported an association between milk allergy resolution and
606 higher levels of the genus *Eubacterium*. At visit 12 months, higher *Bacteroidaceae* were
607 observed in infants with persistent CMA. Higher levels of the phylum Bacteroidetes were
608 reported earlier in infants with persistent CMA [52]. For the group with outgrowth of CMA,
609 higher *Rikenellaceae*, *Actinomycetaceae* and *Eggerthellaceae* were found by the LEfSe
610 analysis at visit 12 months. To the best of our knowledge, these families have not been
611 associated earlier with outgrowth of CMA.
612 When comparing visits within allergy groups, both the LMM and LEfSe analysis showed
613 higher *Ruminococcaceae* at 12 months compared to baseline in the group with outgrowth of
614 CMA, while no significant changes in this family were observed in infants with persistent
615 CMA. Bunyavanich et al. [52] reported increased levels of the genus *Ruminococcus* in infants
616 with resolved CMA compared with persistent CMA. Furthermore, both analyses identified
617 higher *Bacteroidaceae* at 12 months compared to baseline in the group with persistent CMA,
618 while no significant changes were found in the group with outgrowth of CMA. Higher levels of
619 Bacteroidetes were associated earlier with persistent CMA [52].
620 In both allergy groups, lower *Enterobacteriaceae* were observed at 12 months compared to
621 baseline visit by both LMM and LEfSe analysis. Decreased levels of *Enterobacteroidaceae*
622 have been previously reported in infants with CMA that received an amino acid-based
623 formula during a period of 6 months [15].
624 In both allergy groups, the results of LEfSe analysis showed higher relative abundances of
625 the families *Butyricicoccaceae* and *Monoglobaceae* at visit 12 months compared to baseline,
626 which suggests that these families change when infants growing older. Higher

627 *Eubacteriaceae*, *Oscillospiraceae* and *Tannerellaceae* at visit 6 months compared to baseline
628 were observed only in the group with outgrowth of CMA. To the best of our knowledge,
629 these three families have not been associated with milk allergy resolution earlier.
630 Furthermore, the LEfSe analysis showed lower *Enterococcaceae*, *Staphylococcaceae* and
631 *Saccharimonadeceae* at visit 6 months compared to baseline in the group with outgrowth of
632 CMA only. Significantly lower *Enterococcaceae* were reported earlier in healthy infants
633 compared to infants with food sensitization [53]. A systematic review [54] reported contradictory
634 results in the literature for *Staphylococcaceae* in relation to food sensitization (i.e. some
635 studies relate food sensitization with higher *Staphylococcaceae*, while others found lower
636 *Staphylococcaceae*). When comparing visit 12 months with baseline visit, higher relative
637 abundances of several families were only found in the group with outgrowth of CMA:
638 *Christensenellaceae*, *Oscillospiraceae*, *Oscillospirales* family, *Clostridia UCG 014* family,
639 *Acidaminococcaceae*, *Rikenellaceae* and *Lachnospiraceae*. Higher relative abundances of
640 *Rikenellaceae* were also reported in healthy infants compared to infants with food
641 sensitization by Chen et al. [53]. The same study also reported higher levels of
642 *Acidaminococcaceae*. However, the results were not significant. Lower *Micrococcaceae* at
643 visit 12 months than baseline visit were found in the group with outgrowth of CMA. To the
644 best of our knowledge, this family has not been previously related to CMA resolution. In
645 infants who outgrew their CMA, but not in those with persistent CMA, higher relative
646 abundances of *Rikenellaceae*, *Actinomycetaceae* and *Prevotellaceae* at visit 12 months
647 compared to visit 6 months were identified. Higher *Prevotellaceae* were related earlier to
648 CMA resolution [52]. Remarkably, when comparing visit 12 months to baseline, higher
649 *Prevotellaceae* were found in the group with persistent CMA, which contradicts with previous
650 findings [52]. In the group with persistent CMA only, higher *Leuconostocaceae* at 12 months
651 compared to baseline, and higher *Selemonadaceae* at 12 months compared to 6 months
652 were observed. As far as we are aware, these two families have not been related to
653 persistent CMA earlier.
654 A much lower amount of significant results were identified by the metaproteomics analysis at
655 the taxonomy level compared to the analysis of the 16S rRNA gene-based taxonomy. This
656 suggests that changes related to outgrowth of allergy mainly occur at the level of the 16S
657 rRNA gene-based taxonomic classification and less at the protein-based taxonomic
658 classification level. When comparing allergy groups within visits, the only significant result at
659 the family level was higher *Eggerthellaceae* proteins in infants with persistent CMA compared
660 to outgrowth of CMA, which was obtained by the LEfSe analysis. To the best of our
661 knowledge, this result has not been reported earlier.
662 Both LMM and LEfSe analysis showed lower relative abundance of *Enterobacteriaceae* at visit
663 6 months compared to baseline in infants who outgrew their CMA only, while a similar result
664 was found in both allergy groups from the 16S rRNA gene sequencing data when comparing
665 visit 12 months to baseline.
666 Increasing relative abundances of *Coriobacteriaceae* and *Veillonellaceae* proteins over time
667 were found by LEfSe analysis for both allergy groups. For *Coriobacteriaceae*, these changes
668 occurred within another time interval. The change in both allergy groups suggests that

differences in proteins of these two families could be due to the infants that are growing older. Furthermore, higher relative abundances of *Lactobacillaceae* proteins at visit 12 months compared to 6 months were found in the group with outgrowth of CMA only. Kingkaw et al. [17] showed that several *Lactobacillaceae* proteins were uniquely expressed in infants with atopic dermatitis.

The LMM analysis on microbial protein functional classes did not identify any significant differences between allergy groups and visits, suggesting that changes related to outgrowth of CMA occur at the taxonomic level and not at the protein functional level.

To determine differences in human protein classes between allergy groups and visits, LMM analysis was performed. For children that are still allergic to cow's milk at 12 months, significantly lower levels of human immunoglobulins were observed at visit 12 months compared to the baseline visit. This could indicate that these infants are also developing tolerance to CM, but it would take more time before the allergy is resolved. Furthermore, higher relative abundances of proline-rich proteins at 12 months compared to baseline were observed for both allergy groups, which suggests that this finding is related to infants growing older.

The results of the RDA analysis confirm that there is a small contribution of the microbiome to outgrowth of CMA, and that this contribution mainly occurs at the 16S rRNA gene-based taxonomy level.

In summary, we can conclude microbiome differences related to outgrowth of CMA can be mainly identified at the level of the 16S rRNA gene taxonomic level, to a lesser extent at the microbial protein taxonomic level and not at the microbial protein functional level. The overall contribution of the microbiome to outgrowth of CMA is low.

# Data availability

Clinical data are available from Danone Nutricia research upon reasonable request (contact: Harm Wopereis, Danone Nutricia Research, Utrecht, The Netherlands, Harm.Wopereis@danone.com).

The 16S rRNA-gene sequencing data for this study have been deposited in the European Nucleotide Archive (ENA) at EMBL-EBI under accession number PRJEB56782.

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE [55] partner repository with the dataset identifier PXD037190.

The R code for the statistical calculations made in this study has been deposited to a public repository on GitLab (https://git.wur.nl/afsg-microbiology/publication-supplementary-materials/2022-hendrickx-et-al-earlyfit_presto_allergy_study ).

# References


1 Martin, R. *et al.* Early life: gut microbiota and immune development in infancy. *Benef Microbes* **1**, 367-382, doi:10.3920/BM2010.0027 (2010).

2 Tanaka, M. & Nakayama, J. Development of the gut microbiota in infancy and its impact on health in later life. *Allergol Int* **66**, 515-522, doi:10.1016/j.alit.2017.07.010 (2017).

3 Backhed, F. *et al.* Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life. *Cell Host Microbe* **17**, 852, doi:10.1016/j.chom.2015.05.012 (2015).

4 Koenig, J. E. *et al.* Succession of microbial consortia in the developing infant gut microbiome. *Proc Natl Acad Sci U S A* **108 Suppl 1**, 4578-4585, doi:10.1073/pnas.1000081107 (2011).

5 Lee, E. *et al.* Dynamics of Gut Microbiota According to the Delivery Mode in Healthy Korean Infants. *Allergy Asthma Immunol Res* **8**, 471-477, doi:10.4168/aair.2016.8.5.471 (2016).

6 Stefka, A. T. *et al.* Commensal bacteria protect against food allergen sensitization. *Proc Natl Acad Sci U S A* **111**, 13145-13150, doi:10.1073/pnas.1412008111 (2014).

7 Shu, S. A. *et al.* Microbiota and Food Allergy. *Clin Rev Allergy Immunol* **57**, 83-97, doi:10.1007/s12016-018-8723-y (2019).

8 Cukrowska, B., Bierla, J. B., Zakrzewska, M., Klukowski, M. & Maciorkowska, E. The Relationship between the Infant Gut Microbiota and Allergy. The Role of Bifidobacterium breve and Prebiotic Oligosaccharides in the Activation of Anti-Allergic Mechanisms in Early Life. *Nutrients* **12**, doi:10.3390/nu12040946 (2020).

9 Azad, M. B. *et al.* Infant gut microbiota and food sensitization: associations in the first year of life. *Clin Exp Allergy* **45**, 632-643, doi:10.1111/cea.12487 (2015).

10 Chinthrajah, R. S., Hernandez, J. D., Boyd, S. D., Galli, S. J. & Nadeau, K. C. Molecular and cellular mechanisms of food allergy and food tolerance. *J Allergy Clin Immunol* **137**, 984-997, doi:10.1016/j.jaci.2016.02.004 (2016).

11 Nakayama, J. *et al.* Aberrant structures of fecal bacterial community in allergic infants profiled by 16S rRNA gene pyrosequencing. *FEMS Immunol Med Microbiol* **63**, 397-406, doi:10.1111/j.1574-695X.2011.00872.x (2011).

12 Bjorksten, B., Naaber, P., Sepp, E. & Mikelsaar, M. The intestinal microflora in allergic Estonian and Swedish 2-year-old children. *Clin Exp Allergy* **29**, 342-346, doi:10.1046/j.1365-2222.1999.00560.x (1999).

13 Huang, Y. J. *et al.* The microbiome in allergic disease: Current understanding and future opportunities-2017 PRACTALL document of the American Academy of Allergy, Asthma & Immunology and the European Academy of Allergy and Clinical Immunology. *J Allergy Clin Immunol* **139**, 1099-1110, doi:10.1016/j.jaci.2017.02.007 (2017).

14 Penders, J. *et al.* Gut microbiota composition and development of atopic manifestations in infancy: the KOALA Birth Cohort Study. *Gut* **56**, 661-667, doi:10.1136/gut.2006.100164 (2007).

15 Dong, P., Feng, J. J., Yan, D. Y., Lyu, Y. J. & Xu, X. Early-life gut microbiome and cow's milk allergy- a prospective case - control 6-month follow-up study. *Saudi J Biol Sci* **25**, 875-880, doi:10.1016/j.sjbs.2017.11.051 (2018).

16 Kleiner, M. Metaproteomics: Much More than Measuring Gene Expression in Microbial Communities. *mSystems* **4**, doi:10.1128/mSystems.00115-19 (2019).

17 Kingkaw, A. *et al.* Analysis of the infant gut microbiome reveals metabolic functional roles associated with healthy infants and infants with atopic dermatitis using metaproteomics. *PeerJ* **8**, e9988, doi:10.7717/peerj.9988 (2020).

18 Chatchatee, P. *et al.* Tolerance development in cow's milk-allergic infants receiving amino acid-based formula: A randomized controlled trial. *J Allergy Clin Immunol*, doi:10.1016/j.jaci.2021.06.025 (2021).
19 Batth, T. S. *et al.* Protein Aggregation Capture on Microparticles Enables Multipurpose Proteomics Sample Preparation. *Mol Cell Proteomics* **18**, 1027-1035, doi:10.1074/mcp.TIR118.001270 (2019).
20 Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30**, 614-620, doi:10.1093/bioinformatics/btt593 (2014).
21 Schloss, P. D. *et al.* Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* **75**, 7537-7541, doi:10.1128/AEM.01541-09 (2009).
22 Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahe, F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**, e2584, doi:10.7717/peerj.2584 (2016).
23 Haas, B. J. *et al.* Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res* **21**, 494-504, doi:10.1101/gr.112730.110 (2011).
24 Amir, A. *et al.* Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. *Msystems* **2**, e00191-00116 (2017).
25 Bolyen, E. *et al.* Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol* **37**, 852-857, doi:10.1038/s41587-019-0209-9 (2019).
26 Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology* **73**, 5261-5267, doi:10.1128/Aem.00062-07 (2007).
27 Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* **41**, D590-596, doi:10.1093/nar/gks1219 (2013).
28 Johnson, M. *et al.* NCBI BLAST: a better web interface. *Nucleic Acids Res* **36**, W5-9, doi:10.1093/nar/gkn201 (2008).
29 UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res* **49**, D480-D489, doi:10.1093/nar/gkaa1100 (2021).
30 Waterhouse, R. M. *et al.* BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Mol Biol Evol* **35**, 543-548, doi:10.1093/molbev/msx319 (2018).
31 Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150-3152, doi:10.1093/bioinformatics/bts565 (2012).
32 Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* **26**, 1367-1372, doi:10.1038/nbt.1511 (2008).
33 Cox, J. *et al.* Andromeda: a peptide search engine integrated into the MaxQuant environment. *J Proteome Res* **10**, 1794-1805, doi:10.1021/pr101065j (2011).
34 R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing, Vienna, Austria, 2019).
35 Lahti, L. & Shetty, S. microbiome R package. (2012-2019).
36 Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research* **28**, 27-30 (2000).
37 Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 289-300 (1995).

38 Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J. A. & Pawlowsky-Glahn, V. Logratio analysis and compositional distance. *Mathematical Geology* **32**, 271-275 (2000).
39 Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. & Egozcue, J. J. Microbiome datasets are compositional: and this is not optional. *Frontiers in microbiology* **8**, 2224 (2017).
40 Oksanen, J. *et al.* (2020).
41 Segata, N. *et al.* Metagenomic biomarker discovery and explanation. *Genome biology* **12**, 1-18 (2011).
42 Xinias, I., Cassimos, D., Trypsianis, G., Nivatsi, M. & Mavroudi, A. Immediate vs delayed cow's milk protein allergy in terms of tolerance at year 1. *Ann Allergy Asthma Immunol* **123**, 304-306, doi:10.1016/j.anai.2019.06.024 (2019).
43 Faye, O. *et al.* Validation of the Patient-Oriented SCORing for Atopic Dermatitis tool for black skin. *J Eur Acad Dermatol Venereol* **34**, 795-799, doi:10.1111/jdv.15999 (2020).
44 Canani, R. B. *et al.* Lactobacillus rhamnosus GG-supplemented formula expands butyrate-producing bacterial strains in food allergic infants. *Isme J* **10**, 742-750, doi:10.1038/ismej.2015.151 (2016).
45 Barko, P. C., McMichael, M. A., Swanson, K. S. & Williams, D. A. The Gastrointestinal Microbiome: A Review. *J Vet Intern Med* **32**, 9-25, doi:10.1111/jvim.14875 (2018).
46 Niu, J. *et al.* Evolution of the Gut Microbiome in Early Childhood: A Cross-Sectional Study of Chinese Children. *Front Microbiol* **11**, 439, doi:10.3389/fmicb.2020.00439 (2020).
47 Duysburgh, C., Van den Abbeele, P., Krishnan, K., Bayne, T. F. & Marzorati, M. A synbiotic concept containing spore-forming Bacillus strains and a prebiotic fiber blend consistently enhanced metabolic activity by modulation of the gut microbiome in vitro. *Int J Pharm X* **1**, 100021, doi:10.1016/j.ijpx.2019.100021 (2019).
48 Zwittink, R. D. *et al.* Metaproteomics reveals functional differences in intestinal microbiota development of preterm infants. *Mol Cell Proteomics* **16**, 1610-1620, doi:10.1074/mcp.RA117.000102 (2017).
49 De Filippis, F. *et al.* Specific gut microbiome signatures and the associated pro-inflamatory functions are linked to pediatric allergy and acquisition of immune tolerance. *Nat Commun* **12**, 5958, doi:10.1038/s41467-021-26266-z (2021).
50 Bjorksten, B., Sepp, E., Julge, K., Voor, T. & Mikelsaar, M. Allergy development and the intestinal microflora during the first year of life. *J Allergy Clin Immunol* **108**, 516-520, doi:10.1067/mai.2001.118130 (2001).
51 Kourosh, A. *et al.* Fecal microbiome signatures are different in food-allergic children compared to siblings and healthy children. *Pediatr Allergy Immunol* **29**, 545-554, doi:10.1111/pai.12904 (2018).
52 Bunyavanich, S. *et al.* Early-life gut microbiome composition and milk allergy resolution. *J Allergy Clin Immunol* **138**, 1122-1130, doi:10.1016/j.jaci.2016.03.041 (2016).
53 Chen, C. C., Chen, K. J., Kong, M. S., Chang, H. J. & Huang, J. L. Alterations in the gut microbiotas of children with food sensitization in early life. *Pediatr Allergy Immunol* **27**, 254-262, doi:10.1111/pai.12522 (2016).
54 Zimmermann, P., Messina, N., Mohn, W. W., Finlay, B. B. & Curtis, N. Association between the intestinal microbiota and allergic sensitization, eczema, and asthma: A systematic review. *Journal of Allergy and Clinical Immunology* **143**, 467-485, doi:10.1016/j.jaci.2018.09.025 (2019).
55 Vizcaino, J. A. *et al.* 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res* **44**, 11033, doi:10.1093/nar/gkw880 (2016).

## Acknowledgements

Guus Roeselers (Danone Nutricia Research) is gratefully acknowledged for his input in the study design. We thank Heleen de Weerd (Danone Nutricia Research) for pre-processing the 16S rRNA gene amplicon sequencing data.

## Author contributions

DMH performed the statistical analysis and wrote the majority of the manuscript, with input of all co-authors. RA and SB performed the metaproteomics experiments, and wrote the corresponding parts of the manuscript. DMH and SKM developed the proteomics database, SKM wrote the corresponding part of the manuscript. The PRESTO study team generated the clinical data. JML provided the 16S rRNA gene sequencing data and was responsible for the project management. CB designed the study, coordinated the project and participated in all phases of the research and article preparation. All authors contributed to writing the manuscript. All authors have approved the final article.

## Funding

This study was part of the EARLYFIT project (Partnership programme NWO Domain AES-Danone Nutricia Research), funded by the Dutch Research Council (NWO) and Danone Nutricia Research (project number: 16490).

## Competing interests

Jolanda Lambert was an employee of Danone Nutricia Research. The project is part of a partnership programme between NWO-TTW and Danone Nutricia Research. The other authors declare that they have no known conflicts of interest.

**PRESTO study team**

**Pantipa Chatchatee[4], Anna Nowak-Wegrzyn[5,6], Lars Lange[7], Suwat Benjaponpitak[8], Kok Wee Chong[9], Pasuree Sangsupawanich[10], Marleen T. J. van Ampting[3], Manon M. Oude Nijhuis[3], Lucien F. Harthoorn[3], Jane E. Langford[3], Jan Knol[1,3], Karen Knipping[3], Johan Garssen[3,11], Valerie Trendelenburg[12], Robert Pesek[13], Carla M. Davis[14], Antonella Muraro[15], Mich Erlewyn-Lajeunesse[16], Adam T. Fox[17], Louise J. Michaelis[18], Kirsten Beyer[19], Lee Noimark[20], Gary Stiefel[21], Uwe Schauer[22], Eckard Hamelmann[22], Diego Peroni[23], and Attilio Boner[23]**

[4] Pediatric Allergy and Clinical Immunology Research Unit, Division of Allergy and Immunology, Department of Pediatrics, Faculty of Medicine, Chulalongkorn University, King Chulalongkorn Memorial Hospital, Thai Red Cross Society, Bangkok, Thailand

[5] Allergy and Immunology, Department of Pediatrics, New York University Langone Health, New York, US

[6] Department of Pediatrics, Gastroenterology and Nutrition, Collegium Medicum, University of Warmia and Mazury, Olsztyn, Poland

[7]St Marien Hospital,Bonn, Germany

[8]Pediatric Allergy and Immunology Division, Department of Pediatrics, Faculty of Medicine Ramathibodi Hospital, Mahidol University, Bangkok, Thailand

[9]Allergy Service, Department of Paediatric Medicine, KK Women's & Children's Hospital, Singapore

[10]Department of Pediatrics, Faculty of Medicine, Prince of Songkla University, Hat Yai, Thailand

[11]Utrecht Institute for Pharmaceutical Sciences, Utrecht University, Utrecht, The Netherlands

[12] Department of Pediatric Pneumology, Immunology, and Intensive Care Medicine, Charité Universitätsmedizin Berlin, Berlin, Germany

[13]Arkansas Children's Hospital, Little Rock, US

[14]Texas Children's Hospital, Baylor College of Medicine, Houston, US

[15]Food Allergy Referral Centre, Padua University Hospital, Padua, Italy

[16]University Hospitals Southampton, Southampton, United Kingdom

[17]Guy's and St Thomas' NHS Foundation Trust, London, United Kingdom

[18]Great North Children's Hospital, Newcastle Upon Tyne Hospitals NHS Foundation Trust, Newcastle Upon Tyne, United Kingdom

[19]Department of Pediatric Pneumology, Immunology, and Intensive Care Medicine, Charité Universitätsmedizin Berlin, Berlin, Germany

[20]Barts/Royal London Hospital, London, England, United Kingdom

[21]Leicester Royal Infirmary, Leicester, England, United Kingdom

[22]Ruhr-Universitat Bochem im St Josef-Hospital, Bochum, Germany

[23]University Hospital Verona, Verona, Italy