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; clara.belzer@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.

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, and to a lesser extent at the protein-based microbial taxonomy and 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 outgrew their CMA after one year and others did not. In this way, we aim to identify potential microbial factors associated with outgrowth of CMA.

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 outgrowth of CMA.
IgE-mediated CMA was diagnosed as follows as elsewhere [18]. First, infants were considered as sensitized to cow's milk (CM) if their CM-specific serum IgE was higher than 0.1 kU/L and/or a CM skin prick test (SPT) resulted in a wheal size ≥ 3 mm. Next, diagnosis of IgE-mediated CMA was confirmed by an open or double-blind placebo-controlled cow's milk challenge or a history of anaphylaxis reaction to isolated ingestion of CM confirmed by two physicians.
Outgrowth of CMA at visit 12 months was evaluated through a double-blind placebo-controlled food challenge (DBPCFC) with CM powder, followed by an oral fresh milk challenge in case the former was negative, like described earlier [18].
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, outgrowth of CMA 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) outgrew their CMA 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 CMA 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].

Faeces sample collection and storage

Faeces samples for 16S-rRNA gene amplicon sequencing and metaproteomics were collected in 30-mL stool containers (Greiner 443102, Merck) at baseline, 6 months and 12 months. Aliquots were stored in 1.5 ml eppendorf tubes. Collection of stools took place between 2013 – 2018, and as a consequence the time gap between collection and analysis differs per sample. Faeces samples collected at home were immediately stored in home freezers of the parents and then transported with ice-packs to the hospitals within three months, where they were stored at -80 °C. Thereafter, faeces samples were transported from the hospitals to Danone Nutricia Research (the Netherlands) on dry-ice and stored at -80 °C until analysis. Transport of samples to LifeSequencing S.L. (Valencia, Spain) for 16S-rRNA gene amplicon sequencing, and to Wageningen University (the Netherlands) for metaproteomics was done in the same way. At Wageningen University, the samples for metaproteomics were stored at -80 °C until sample preparation. Peptide samples were stored at -20 °C until they were measured by nLC-MS/MS.

DNA extraction for 16S-rRNA gene amplicon sequencing

DNA was extracted from faeces samples with QIAmp DNA Stool Mini Kit (Qiagen, Venlo, the Netherlands) according to the protocol of the manufacturer, but incorporating two extra bead-beating steps as described in Mischke et al[19].

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).

The number of sequences (reads) per sample varied between 16830 and 66760. In total 713981 amplicon sequence variants were analyzed, of which 13309 were retained after applying the filtering steps described below.

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

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

187

188 ## Metaproteomics nLC-MS/MS

189 Five microliter of peptide sample (defrozen and centrifuged at 12.000 * g for 30 min) was
190 loaded directly onto a 0.10 * 250 mm ReproSil-Pur 120 C18-AQ 1.9 µm beads analytical
191 column (prepared in-house) at a constant pressure of 825 bar (flow rate of circa 650
192 nL/min) with 1 ml/l HCOOH in water and eluted at a flow of 0.5 µl/min with a 50 min linear
193 gradient from 9% to 34% acetonitril in water with 1 ml/l formic acid with a Thermo EASY
194 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 [21], 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 [22], and reads with < 2 reads over all samples were removed. Next, VSEARCH [23] together with the RDP gold database [24] were used to eliminate chimeras. Filtering of reads including PhiX and Adapter sequences (as defined in Deblur [25]) was performed using QIIME2 [26]. For each amplicon sequence variant (ASV), taxonomy was assigned at the genus level using the Ribosomal Database Project (RDP) [27] classifier against the SILVA 138 reference database [28], resulting in a classification of the ASVs into 173 genera.

Database construction

ASVs were identified at the species level with NCBI nucleotide BLAST [29], 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 each ASV, the NCBI sequence with the highest percent identity was selected, without using any threshold. In case multiple NCBI sequences fulfilled this condition, they were all retained. The selected NCBI sequences were aggregated at the species level. For the resulting 327 species the non-redundant proteomes were obtained from UniProtKB [30]. 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) [30] and Benchmarking

Universal Single-Copy Ortholog (BUSCO) [31] 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 [32] 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 2.0.3.0 [33] using the "Specific Trypsin/P" Digestion mode with maximally 2 missed cleavages and further default settings for the Andromeda search engine [34] (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 2705 protein groups, of which 2481 were from bacteria.

Statistical analysis of 16S rRNA gene sequencing and metaproteomics data

All analyses (except LEfSe, see below) were performed in R version 3.6.1 [35]. 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 [36] version 1.8.0. For the metaproteomics data, relative

273 abundances were calculated from the intensity Based Absolute Quantifcation (iBAQ)
274 intensities.
275 As 16S rRNA gene-based taxonomic classification at genus level included a considerable
276 fraction of unclassified *Lachnospiraceae* (Figure S1), statistical analysis was performed at
277 the family level. Core microbiota were determined at the family level using the R microbiome
278 package. Core taxa were defined as taxa that have a relative abundance higher than 1% in
279 at least 50% of the 16S rRNA sequencing samples or at least 50% of the metaproteomics
280 samples.
281 Bacterial proteins were functionally annotated by assigning protein identifiers to Kyoto
282 Encyclopedia of Genes and Genomes (KEGG) [37] orthology (KO) identifiers, and classifying
283 the KO identifiers using KEGG Brite hierarchy level C.
284 To compare relative abundances of core taxa, top 10 microbial protein functional classes
285 and top 10 human protein classes within allergy groups and between visits, the data were
286 normalized by centred log ratio (CLR) transform and a linear mixed model (LMM) analysis,
287 with outgrowth of CMA, visit and outgrowth of CMA x visit as fixed effects and subject as
288 random effect, was performed for each core taxon or protein class. The interaction term
289 outgrowth of CMA x visit was included as we expected that visits compare differently at the
290 level of outgrowth of CMA. The LMM was also used to determine differences in core taxa
291 between allergy groups within each visit. P-values for pairwise comparisons were calculated
292 from the model, and adjusted for multiple testing using the Benjamini-Hochberg procedure
293 [38]. To study the effect of the wide age range of the infants when entering the study (3-13
294 months), we conducted an additional LMM analysis adding age as fixed effect. As adding
295 more variables to the model can lead to overfitting, we also compared the model with and
296 without age using Akaike Information Criterion (AIC).
297 To relate variability in the microbiome to environmental and clinical factors, redundancy
298 analysis (RDA) was performed. Aitchison distance [39] (Euclidian distance on CLR transformed
299 data), which has been proposed as a suitable distance metric for compositional data [40], was
300 used for RDA. RDA analysis was performed with the R vegan package [41] version 2.5.-6,
301 using stepwise forward and backward model selection where a variable is added when its p-
302 value is smaller than or equal to 0.05, and removed when its p-value is larger than 0.1.
303 Furthermore, partial RDA with outgrowth of CMA as explanatory variable and the remainder
304 of the environmental variables as covariates was performed to determine the proportion of
305 variance in the microbiome explained by outgrowth of CMA. A similar RDA and partial RDA
306 analysis was performed to relate variability in the microbial proteome and human proteins to
307 environmental factors. Furthermore, another RDA was conducted to relate variability in the
308 proteome and protein-based microbial composition to human proteins.
309 Spearman correlation between 16S gene- and protein-based taxonomic classification, and
310 between human and microbial proteins, were calculated. P-values were determined with
311 Monte Carlo permutation (10 000 permutations), and correlations with a p-value below 0.05
312 were considered significant.
313 Discriminative microbial features (taxa, microbial protein functional classes, human protein
314 classes) 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 [42], 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.
LEfSe applies to relative abundances and focuses on effect size for selecting discriminative features. For comparison, we also included a method that applies to clr-transformed abundances and focuses on adjusted p-values rather than effect size. This method consists of fitting a linear mixed model (LMM) to all families or protein classes (instead of only core taxa or top 10 classes), using outgrowth of CMA, visit and outgrowth of CMA x visit as fixed effects and subject as random effect.

Results

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).

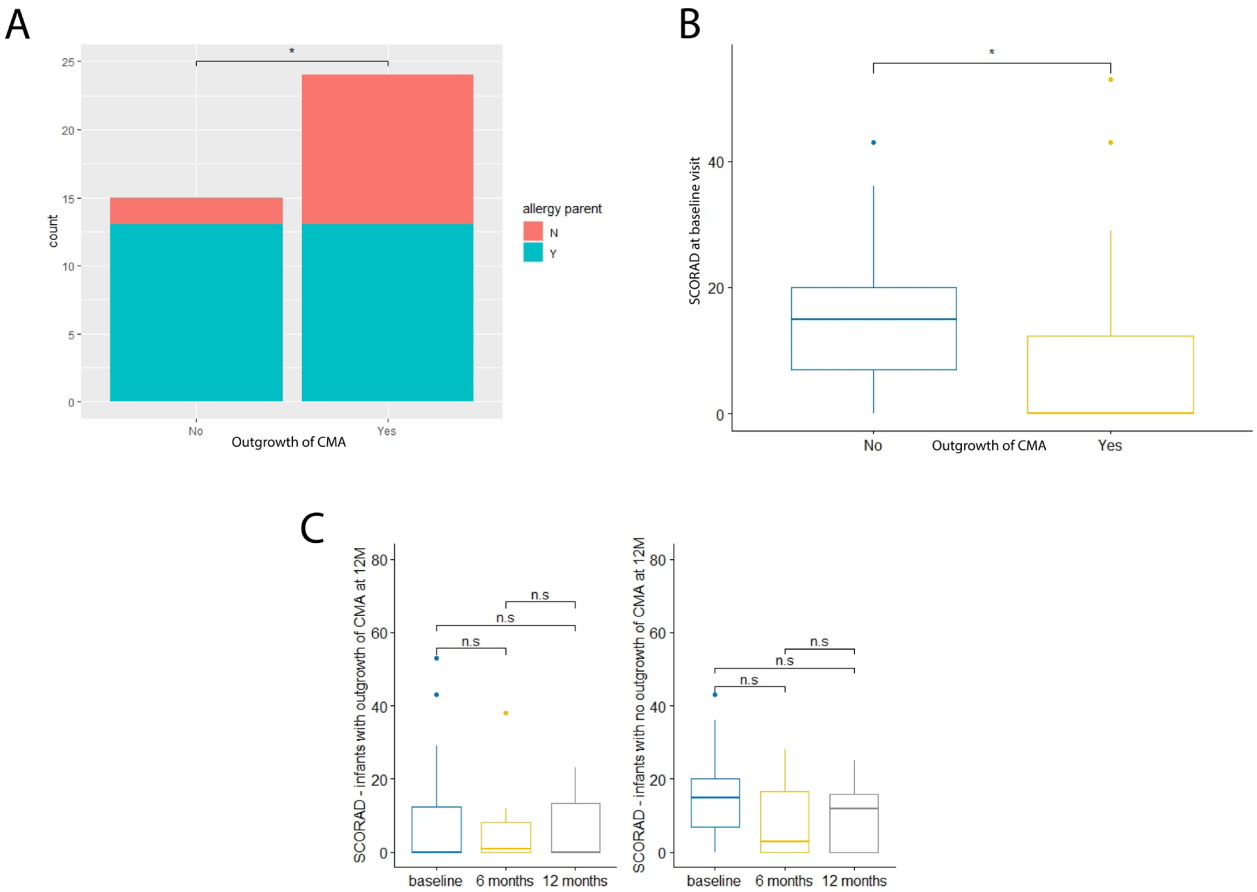

Figure 1. (A) Stacked bar plot for parental allergy. No = no outgrowth of CMA at visit 12 months; Yes = outgrowth of CMA at visit 12 months. (B) Boxplot for SCORAD at baseline visit. No = no outgrowth of CMA at visit 12 months; Yes = outgrowth of CMA at visit 12 months. (C) Boxplots for SCORAD over time. 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 those that outgrew their CMA or those that did not at visit 12 months. The different visits were also analysed separately. For all groups, more than 40% 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).

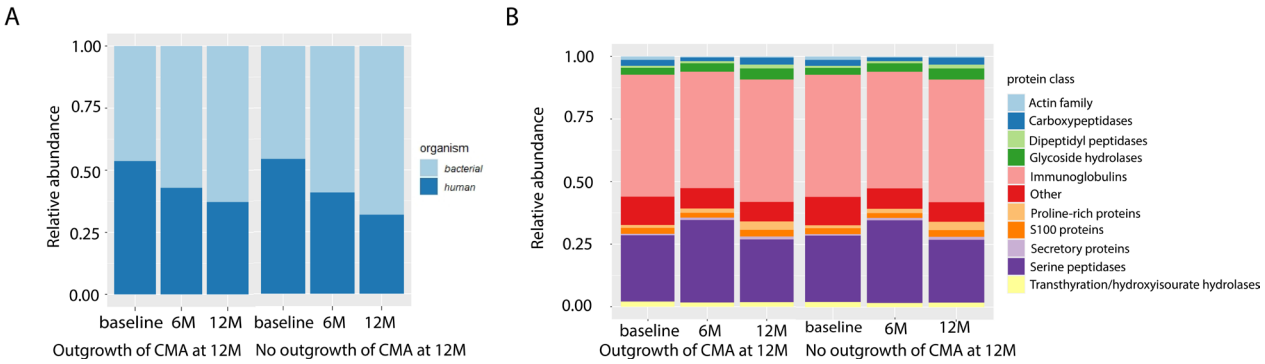

Figure 2. Ratio human/bacterial proteins (A) and human proteins composition (B) at each visit for the group that outgrew their CMA at visit 12 months (12M) and the group that did not.

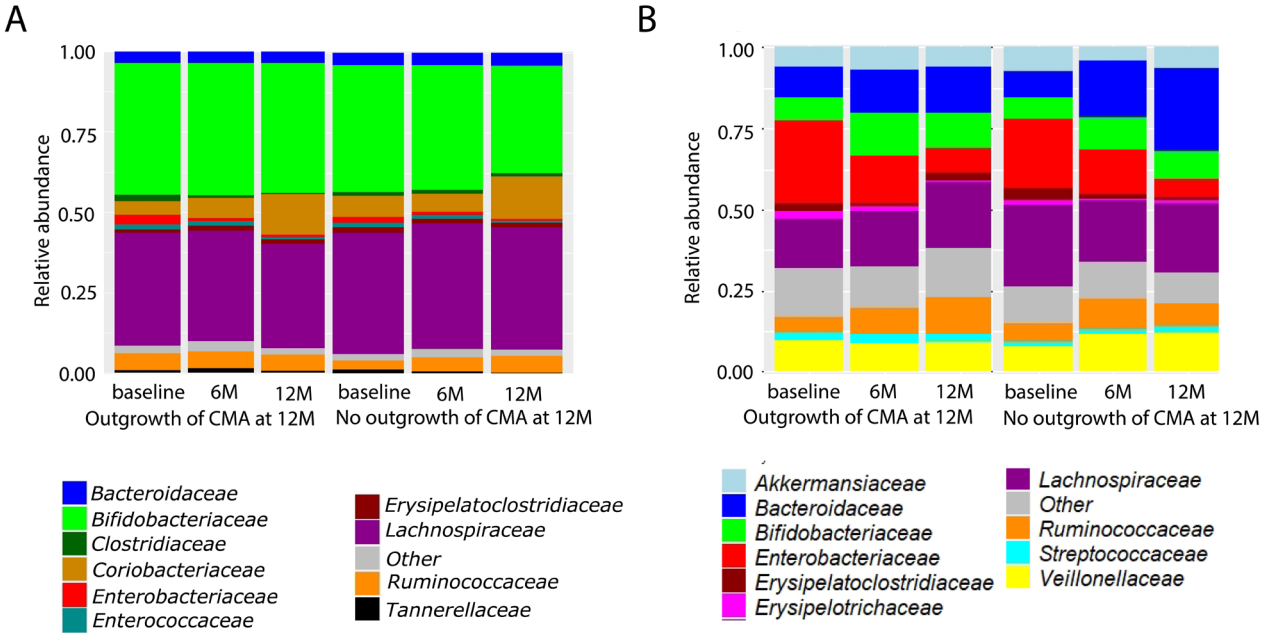

Figure 3. Protein-based microbial taxonomic profiles (family level) (A) and 16S rRNA gene-based taxonomic profiles (B) at each visit for the group that outgrew their CMA at visit 12 months (12M) and the group that did not.

Differences and similarities between 16S rRNA gene-based and protein-based microbial 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 (outgrowth of CMA versus no outgrowth of CMA at visit 12 months) and visits (Figures 3A and B).

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

374 The top 10 most abundant families have 6 taxa in common between 16S rRNA gene-based
375 and protein-based microbial taxonomic profiles. They are: *Bacteroidaceae*,
376 *Bifidobacteriaceae*, *Enterobacteriaceae*, *Erysipelatoclostridiaceae*, *Lachnospiraceae* and
377 *Ruminococcaceae* (Figure 3A and B). *Clostridiaceae*, *Coriobacteriaceae*, *Enterococcaceae* and
378 *Tannerellaceae* were only among the top 10 taxa of protein-based microbial taxonomic
379 profiles (Figure 3A). Taxa that were only among the top 10 in 16S rRNA gene-based profiles
380 were *Akkermansiaceae*, *Erysipelotrichaceae* and *Streptococcaceae* and *Veillonellaceae*
381 (Figure 3B). Both 16S rRNA-based and metaproteomics based taxonomic profiles show high
382 inter-individual variation (Figures S2 and S3).

383

384 ***Correlation core taxa between 16S rRNA gene-based and protein-based microbial***
385 ***profiles***

386 Core taxa, defined as taxa that have a relative abundance higher than 1% in at least 50% of
387 the 16S rRNA gene sequencing or metaproteomics samples included *Bifidobacteriaceae*,
388 *Coriobacteriaceae*, *Bacteroidaceae*, *Lachnospiraceae*, *Ruminococcaceae*, *Veillonellaceae* and
389 *Enterobacteriaceae*. *Bifidobacteriaceae*, *Bacteroidaceae*, *Lachnospiraceae* and
390 *Ruminococcaceae* fulfil these conditions for both 16S rRNA gene sequencing and
391 metaproteomics samples, *Veillonellaceae* and *Enterobacteriaceae* only for 16S rRNA gene
392 sequencing and *Coriobacteriaceae* only for metaproteomics. The relative abundance of
393 *Bifidobacteriaceae*, *Bacteroidaceae*, *Lachnospiraceae*, *Ruminococcaceae* and
394 *Enterobacteriaceae* correlated significantly (p-value ≤ 0.05) between 16S rRNA-based and
395 protein-based microbial profiles (Table 1). However, apart from *Bifidobacteriaceae*, the
396 correlations were low.

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

Family	**Spearman correlation**	**p-value**
Bifidobacteriaceae	0.795	**< 1e-04**
Coriobacteriaceae	0.142	0.1194
Bacteroidaceae	0.284	**0.0024**
Lachnospiraceae	0.539	**< 1e-04**
Ruminococcaceae	0.518	**< 1e-04**
Veillonellaceae	0.073	0.4268
Enterobacteriaceae	0.386	**< 1e-04**

400

401

402

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 and LMM analysis on the whole data set were performed. For all core taxa, the model without age as fixed effect shows the best fit to the data (lowest AIC)(Table S4). 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 both LMM analyses on the core taxa showed significantly higher relative abundances of *Lachnospiraceae* at baseline in infants with persistent CMA compared to infants who outgrew their CMA (Table S5). Furthermore, according the LEfSe analysis the families *Carnobacteriaceae*, *Lachnospiraceae* and *Peptostreptococcales-Tissierellales_fa* 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).

The LMM analysis based on all taxa at the family level found no significant differences between allergy classes within visits after adjusting for multiple testing (Table S6). However, with the exception of *Eggerthellaceae* at 12 months, all discriminative features found by LEfSe were in the top 10 of LMM ordered by unadjusted p-value (Table S6). This indicates concordance between the LMM and LEfSe results, and suggests that the taxa obtained by LEfSe could have been significant when the LMM analysis could have been repeated in a larger cohort.

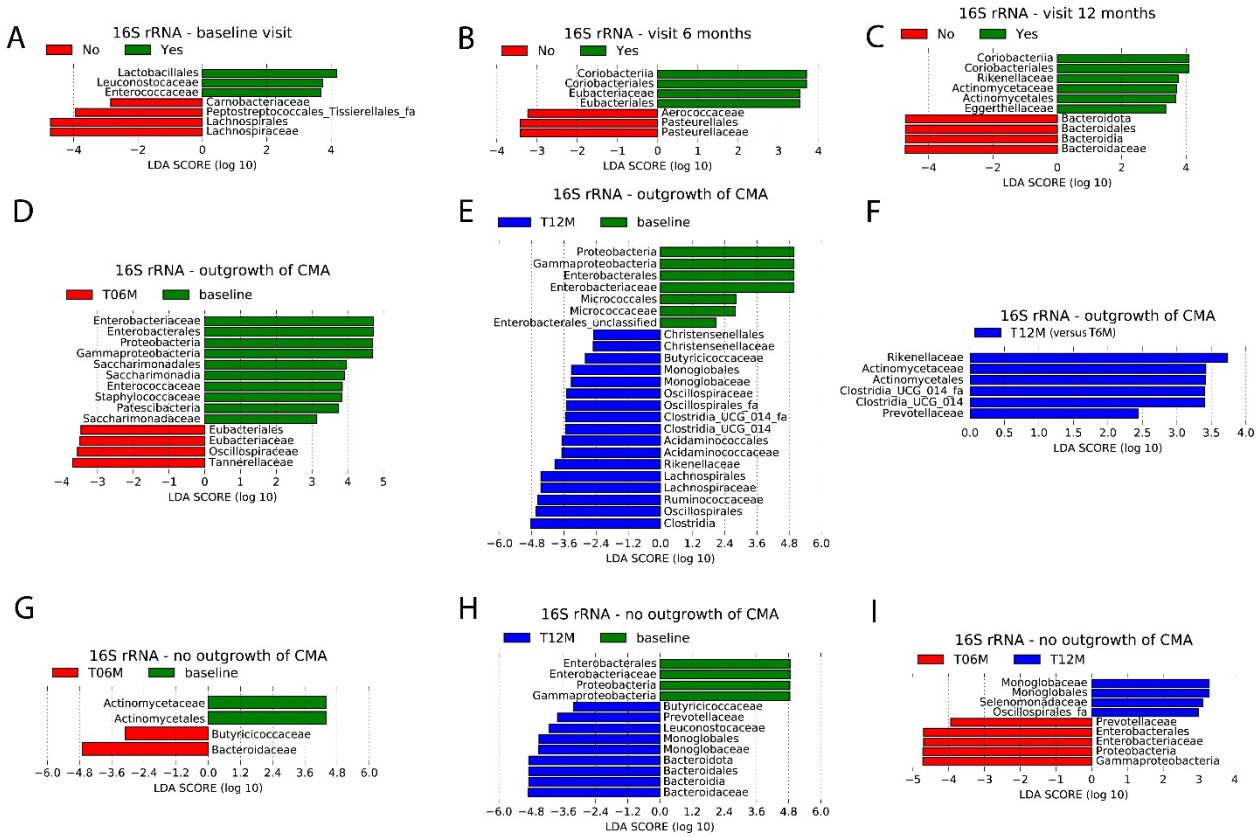

Figure 4. LEfSe analysis of 16S rRNA-based relative abundances at the family and all higher taxonomic levels using an alpha value of 0.05 for the factorial Kruskal-Wallis test among classes and a threshold of 2.0 on the logarithmic LDA score for discriminative features. (A-C) Using outgrowth of CMA at visit 12 months as class, Yes: outgrowth of CMA at 12 months; no: no outgrowth of CMA 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 S7 and Figure S4). In the group with persistent CMA, relative abundances of *Bacteroidaceae* were significantly higher at visit 12 months compared to baseline visit (Table S7 and Figure S4). 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 baseline visit, as well as at visit 6 months compared to baseline visit (Table S7 and Figure S4). In the group with persistent CMA, *Enterobacteriaceae* were only lower at visit 12 months compared to baseline (Table S7 and Figure S4). When including age as fixed effect in the LMM, we did not observe these effects.

In the group with outgrowth of CMA, the LEfSe analysis showed higher relative abundances of the families *Eubacteriaceae*, *Oscillospiraceae* and *Tannerellaceae*, and lower relative abundances of *Enterobacteriaceae*, *Enterococcaceae*, *Staphylococcaceae* and *Saccharimonadaceae* at visit 6 months compared to baseline (Figure 4D). Higher relative abundances of the families *Christensenellaceae*, *Butyricicoccaceae*, *Monoglobaceae*, *Oscillospiraceae*, *Oscillospirales_fa*, *Clostridia UCG 014* family, *Acidaminococcaceae*, *Rikenellaceae*, *Lachnospiraceae* and *Ruminococcaceae*, and lower relative abundances of *Enterobacteriaceae* and *Micrococcaceae* were observed at visit 12 months compared to baseline visit (Figure 4E). Higher relative abundances of the families *Rikenellaceae*, *Actinomycetaceae* and *Prevotellaceae* were observed at visit 12 months compared to visit 6 months (Figure 4F). The majority of families found by LEfSe in the comparisons 6 months versus baseline and 12 months versus baseline were also significant for the LMM analysis applied to all families (Table S8). For the comparison 12 months versus 6 months, the majority of the families found by LEfSe belongs to the top 10 ordered by unadjusted p-value found for the LMM analysis based on all families (Table S8). This suggests that these taxa could have been significant when the analysis could have been conducted in a larger cohort.

At the family level, LEfSe analysis for the group with persistent CMA showed significantly higher relative abundances of *Butyricicoccaceae* and *Bacteroidaceae*, and significantly lower relative abundances of *Actinomycetaceae* at visit 6 months compared to baseline (Figure 4G). Higher relative abundances of the families *Butyricicoccaceae*, *Prevotellaceae*, *Leuconostocaceae*, *Monoglobaceae* and *Bacteroidaceae*, as well as lower relative abundances of *Enterobacteriaceae* were observed at visit 12 months compared to baseline (Figure 4H). Moreover, higher relative abundance levels of the families *Monoglobaceae* and *Selenomonadaceae*, as well as lower relative abundance levels of *Prevotellaceae* and *Enterobacteriaceae* were observed at visit 12 months compared to visit 6 months (Figure 4I). The majority of the results obtained by LEfSe for this allergy group are not significant, but belong to the top 10 ordered by unadjusted p-value (Table S9).

Significant differences between allergy groups (outgrowth versus no outgrowth of CMA) in protein-based microbial relative abundance levels

Linear Mixed Model (LMM) and LEfSe analysis were performed to determine microbial variables most likely explaining differences between allergy groups in the metaproteomics data. Similar as for the models based on 16S rRNA gene-based relative abundance levels, the models without age as fixed effect show the best goodness of fit (lowest AIC)(Table S10). Below, differences at the family level are reported in detail, for differences at the higher taxonomy levels we refer to Figure 5.

Significant protein-based microbial differences between between allergy groups within each visit

Both LMM analyses did not identify any significant differences in core taxa between allergy groups within visits (Table S11). The results of the LEfSe analysis show higher relative abundances of *Eggerthellaceae* proteins at baseline and higher relative abundances of *Veillonelleceae* proteins at visit 6 months in infants with persistent CMA compared to infants who outgrew their CMA (Figure 5A and B). For visit 12 months, no significant differences between allergy groups were observed.

The LMM analysis based on all taxa at the family level found no significant differences between allergy classes within visits after adjusting for multiple testing (Table S12). However, *Eggerthellaceae* has the lowest uncorrected p-value for baseline visit, and *Veillonellaceae* has the second lowest uncorrected p-value for 6 months (Table S12). This suggest that these taxa could have been found significant by LMM analysis in a larger cohort with more statistical power.

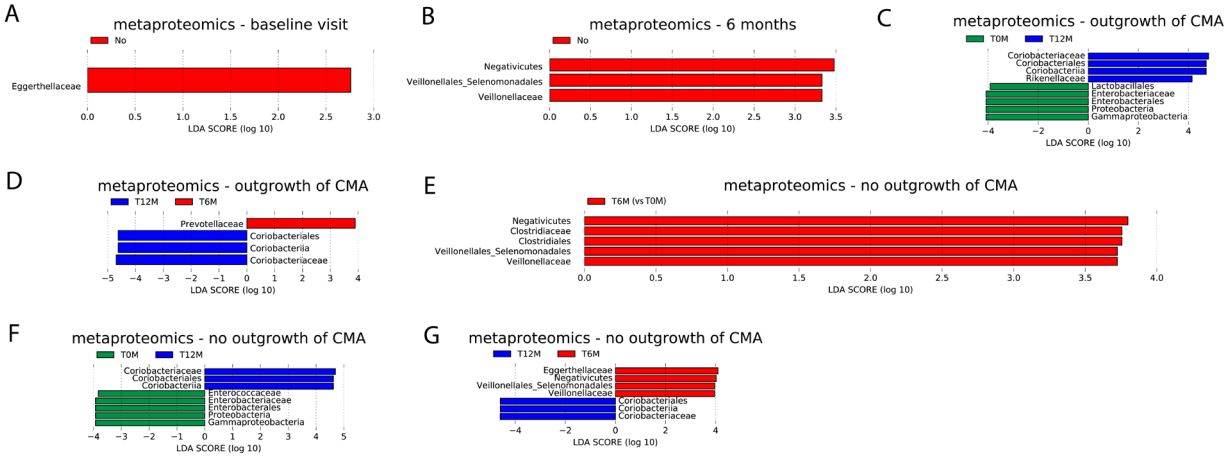

Figure 5. LEfSe analysis of protein-based microbial relative abundances at the family and all higher taxonomic levels using an alpha value of 0.05 for the factorial Kruskal-Wallis test among classes and a threshold of 2.0 on the logarithmic LDA score for discriminative features. (A-B) Using outgrowth of CMA at visit 12 months as class, Yes: outgrowth of CMA at 12 months; no: no outgrowth of CMA at 12 months. Plot of discriminative features at (A) baseline visit; (B) visit 6 months. (C-D) Pairwise comparison between visits within the group with outgrowth of CMA, using visit as class. (C) visit 12 months (T12M) versus baseline (T0M); (D) visit 12 months (T12M) versus visit 6 months (T6M). (E-G) Pairwise comparison between visits within the group with persistent CMA, using visit as class. (E) visit 6 months (T6M) versus baseline (T0M); (F) visit 12 months (T12M) versus baseline (T0M); (G) visit 12 months (T12M) versus visit 6 months (T6M).

Significant protein-based microbial 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 12 months compared to the baseline visit (Table S13 and Figure S5). This effect was not observed when repeating the LMM analysis including age as fixed effect. No significant differences between visits were obtained by the LMM analysis in the group with persistent CMA.
In the group with outgrowth of CMA, significantly higher relative abundances of *Coriobacteriaceae* and *Rikenellaceae* 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 significantly lower relative abundances of *Prevotellaceae* proteins were identified at visit 12 months compared to visit 6 months (Figure 5D).
The results of the LEfSe analysis for the group who outgrew their CMA were all among the top 5 features from LMM analysis ordered by unadjusted p-value (Table S14).
The LEfSe analysis comparing visits within the group with persistent CMA identified significantly higher relative abundances of *Clostridiaceae* and *Veillonellaceae* proteins at visit 6 months compared to baseline (Figure 5E). Significantly higher relative abundances of *Coriobacteriaceae* and significantly lower relative abundances of *Enterococcaceae* and *Enterobacteriaceae* proteins were observed at visit 12 months compared to the baseline visit (Figure 5F). Moreover, significantly higher relative abundances of *Coriobacteriaceae* and significantly lower relative abundances of *Eggerthellaceae* and *Veillonellaceae* proteins were observed at visit 12 months compared to visit 6 months (Figure 5G).
The results of the LEfSe analysis for the group who did not outgrew their CMA were all among the top 10 features from LMM analysis ordered by unadjusted p-value (Table S15).

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

With the exception of the models for pentose and glucoronate interconversions, better goodness of fit (lower AIC) was obtained when not including age as fixed effect in the linear mixed models (LMM)(Table S16).

Significant differences of microbial protein functional classes between allergy groups within each visit

The results of LMM analysis, applied to the top 10 microbial protein functional classes, showed a significantly lower relative abundance of glycolysis / gluconeogenesis and pentose and glucuronate interconversions proteins, and a significantly higher relative abundance of RNA degradation proteins in the group with outgrowth of CMA compared to the group with persistent CMA at baseline visit (Table S17). When including age as fixed effect, only

564 glycolysis / gluconeogenesis showed still significant differences between the two allergy
565 groups, while the other effects were only marginally significant (adjusted p-value lower than
566 0.1). Lower relative abundance of pentose and glucuronate interconversions proteins in the
567 group who outgrew their CMA than in the other group were also found by LEfSe (Figure 6A).
568 Moreover, the results of LEfSe also showed several other protein functional classes with
569 lower relative abundance in the group who outgrew their CMA: Nitrogen metabolism;
570 cysteine and methionine metabolism; valine, leucine and isoleucine degradation, inositol
571 phosphate metabolism, pyrimidine metabolism, selenocompound metabolism and
572 streptomycin biosynthesis. Relative abundances of pyrimidine metabolism were also lower in
573 the group who outgrew their CMA at visit 6 months (Figure 6B). Moreover, at 6 months, the
574 group who outgrew their CMA showed higher relative abundances of proteins from amino
575 sugar and nucleotide sugar metabolism, and from glycine, serine and threonine metabolism
576 in the group who outgrew their CMA than in the other group (Figure 6B). At visit 12 months,
577 relative abundances of proteins from starch and sucrose metabolism were lower in the group
578 who outgrew their CMA than in the other group (Figure 6C).

579 Apart from a few exceptions (nitrogen metabolism at baseline, and cysteine and methionine
580 metabolism at baseline), all results of LEfSe were also reported in the top 10 based on
581 unadjusted p-value from the LMM analysis on all functional classes (Table S18)

582 ***Significant differences of microbial protein functional classes between visits within***
583 ***allergy groups***

584 In the group with outgrowth of CMA, but not in the group with persistent CMA, significantly
585 higher relative abundances of glycolysis / gluconeogenesis proteins and significantly lower
586 relative abundances of RNA degradation proteins were found at the later visits (6M and 12M)
587 compared to baseline (Table S19). At visit 12 months only, a significantly higher relative
588 abundance of pentose and glucuronate interconvertions proteins was found compared to
589 baseline in the group with outgrowth of CMA, but not in the group with persistent CMA
590 (Table S19). After including age as fixed effect, significant differences were only observed
591 between baseline and 12 months for glycolysis / gluconeogenesis and pentose and
592 glucuronate interconversions. The result for pentose and glucuronate interconversions was
593 also confirmed by LEfSe.

594 The results of LEfSe showed several other protein functional classes that changed over visit
595 in the group who outgrew their CMA, but not in the other group. Relative abundance of the
596 following protein functional classes were significantly higher at visits 6 months and 12
597 months than at baseline: Selenocompound metabolism, aminoacyl tRNA biosynthesis and
598 cysteine and methionine metabolism were higher at visits 6 months and 12 months than at
599 baseline (Figures 6D and E). Furthermore, amino sugar and nucleotide sugar metabolism, as
600 well as galactose metabolism increased between baseline and 6 months in this group, while
601 ABC transporters decreased (Figure 6D). Amino sugar and nucleotide sugar metabolism also
602 increased between 6 months and 12 months (Figure 6F). Moreover, several protein
603 functional classes increased between baseline and 12 months (Figure 6E): ribosome,

pyruvate metabolism, posphyrin metabolism, nitrogen metabolism, fructose and mannose metabolism, selenocompound metabolism, beta alanine metabolism, fatty acid degradation, propanoate metabolism, glycerolipid metabolism and streptomycin biosynthesis. For sphingolipid metabolism and other glycan degradation, a decrease between baseline and 12 months was observed (Figure 6E).

For the group who outgrew their CMA, the overlap between the LEfSe and the top 10 based on uncorrected p-value from LMM is only moderate (exception: 12 months versus 6 months, were the two functional classes found by LEfSe are in top 3 from LMM based on uncorrected p-value)(Table S20).

LEfSe results also show several protein functional classes that only change over visits in the group which did not outgrew their CMA: gluthatione metabolism (increase between baseline and 6 months (Figure 6G), decrease between 6 months and 12 months(Figure 6I)), cysteine biosynthesis (increase between baseline and 12 months (Figure 6H)), pantenoate CoA biosynthesis (increase between 6 months and 12 months (Figure 6I)), starch and sucrose metabolism (increase between 6 months and 12 months (Figure 6I)).

For the group who did not outgrew their CMA, all protein functional classes found by LEfSe, except one (oxidative phosphorylation) are in the top 10 of LMM based on unadjusted p-value (Table S21).

Finally, also protein classes that changed between visits in both allergy groups were found: oxidative phosphorylation (increase between baseline and 12 months (Figure 6E and H)), tryptophan metabolism (decrease between baseline and 12 months (Figure 6E and H)) and flagellar assembly (increase between 6 months and 12 months (Figure 6F and I)).

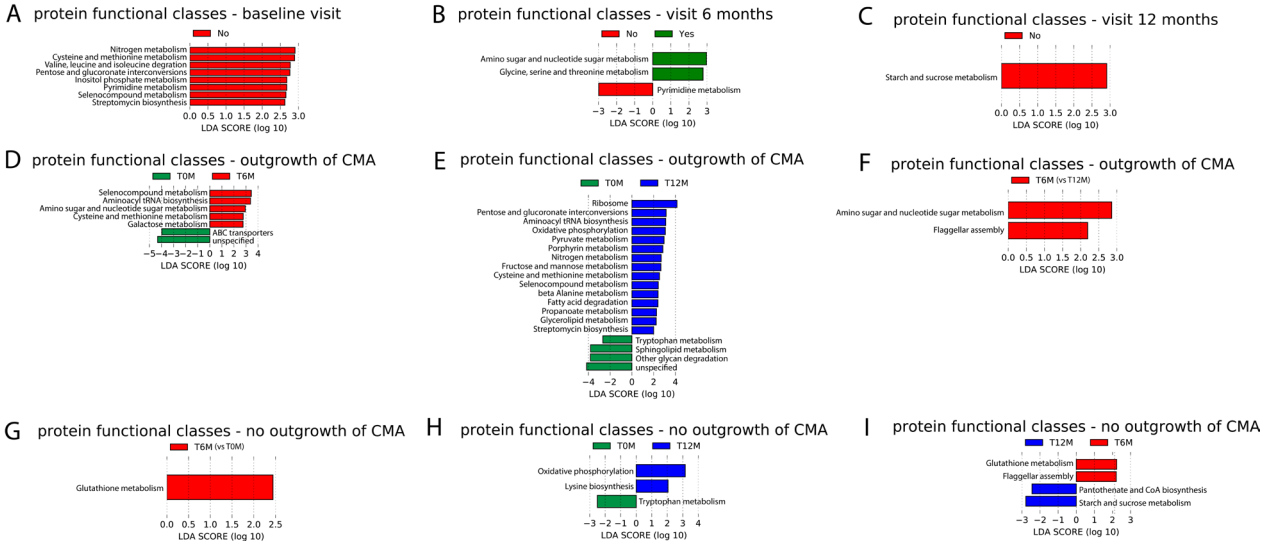

Figure 6. LEfSe analysis of microbial protein functional classes (KEGG Brite level c) using an alpha value of 0.05 for the factorial Kruskal-Wallis test among classes and a threshold of 2.0 on the logarithmic LDA score for discriminative features. (A-C) Using outgrowth of CMA at visit 12 months as class, Yes: outgrowth of CMA at 12 months; no: no outgrowth of CMA at

12 months. Plot of discriminative features at (A) baseline visit; (B) visit 6 months; (C) visit 12 months. (D-F) Pairwise comparison between visits within the group with outgrowth of CMA, using visit as class. (D) visit 6 months (T6M) versus baseline (T0M); (E) visit 12 months (T12M) versus baseline (T0M); (F) visit 12 months (T12M) versus visit 6 months (T6M). (G-I) Pairwise comparison between visits within the group with persistent CMA, using visit as class. (G) visit 6 months (T6M) versus baseline (T0M); (H) visit 12 months (T12M) versus baseline (T0M); (I) visit 12 months (T12M) versus visit 6 months (T6M).

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

Significant differences of human protein classes between allergy groups within each visit

To determine if human protein classes could explain differences between allergy groups, Linear Mixed Model (LMM) analysis was applied. For all human protein classes, the models without age as fixed effect show better goodness of fit (lower AIC)(Table S22). The results show no significant differences between allergy groups within visits (Table S23).

The results of LEfSe analysis showed higher relative abundances of S100 proteins and alkaline phosphatases at baseline in the group who outgrew their CMA compared to the other group (Figure 7A). In the group who outgrew their CMA, lower relative abundances of the actin family and exosomal proteins were found at visit 6 months (Figure 7B). Furthermore, at visit 12 months, higher relative abundances of serpins and lower relative abundances of P-ATPases were found in the group who outgrew their CMA (Figure 7C).

The results of the LEfSe analysis were not significant for the LMM analysis on all human proteins, but all belonged to the top 10 based on unadjusted p-value (Table S24).

Significant differences of human protein classes between visits within allergy groups

According to the LMM results on the top 10 human protein classes, relative abundances of proline-rich proteins (4 protein groups) are significantly higher at visit 12 months compared to the baseline visit in the group with outgrowth of CMA, but not in the group with persistent CMA (Table S25). In the group with outgrowth of CMA only, the relative abundance of S100 proteins (4 protein groups) are significantly lower at visit 6 months than at baseline (Table S25). Moreover, relative abundances of immunoglobulins (37 protein groups) are significantly lower at visit 12 months than baseline in infants with persistent CMA, but not in infants who outgrew their CMA (Table S25). These significant differences between visits were not observed when including age as fixed effect.

668 LEfSe results show higher relative abundances of S100 proteins and alkaline phosphatases in
669 the group who outgrew their CMA at baseline than in the other group (Figure 7A). At visit 6
670 months, lower relative abundances of exosomal proteins and proteins from the actin family
671 were obtained for the group who outgrew their CMA compared to the other group (Figure
672 7B). At visit 12 months, the group who outgrew their CMA had higher serpins and lower P-
673 ATpases than the other group (Figure 7C).

674 Results of LEfSe analysis showed human protein classes that changed over visit in the group
675 who outgrew their CMA, but not in the other group: P-ATPases (increase between baseline
676 and 6 months, Figure 7D); cell surface glycoproteins, S100 proteins and aminopeptidases,
677 transferases (decrease between baseline and 6 months, Figure 7D); dipeptidyl peptidases
678 and serpins (increase between baseline and 12 months and between 6 months and 12
679 months, Figures 7E and F); titin (increase between baseline and 12 months and decrease
680 between 6 months and 12 months, Figures 7E and F); ubiquitins and fused gene family
681 (increase between baseline and 12 months, Figure 7E); carboxypeptidases, superoxide
682 dismutases, binding proteins, ferritins and alkaline phosphatases (decrease between 6
683 months and 12 months, Figure 7E); triose phosphate isomerases, plakins, cadherins and
684 enolases (increase between 6 months and 12 months, Figure 7E).

685 For 6 months versus baseline, and 12 months versus baseline, the majority of the LEfSe
686 results are in the top 10 from the LMM analysis based on unadjusted p-value (Table S26).
687 For 12 months versus 6 months, there is only a moderate overlap between LEfSe resuls and
688 the top 10 list from LMM based on unadjusted p-value (Table S26).

689 LEfSe results also show human protein classes that only change over visits in the group
690 which did not outgrow their CMA: actin family, CUB domain containing proteins (increase
691 between baseline and 6 months and between baseline and 12 months, Figure 7G and H);
692 exosomal proteins (increase between baseline and 6 months, decrease between 6 months
693 and 12 months, Figure 7G and I); triose phosphate isomerases, cell adhesion proteins and
694 actin-binding proteins (increase between baseline and 6 months, Figure 7G); submaxillary
695 gland androgen-regulated proteins, neutral alkaline-ceramides and heme proteins (increase
696 between baseline and 12 months, Figure 7H); amylases (increase between baseline and 12
697 months and between 6 months and 12 months, Figure 7H and I); pyridine nucleotide
698 disulphate reductases (increase between 6 months and 12 months, Figure 7I); proline-rich
699 proteins (increase between baseline and 6 months and between 6 months and 12 months,
700 Figures 7 G and I); serpins and filaggrins (decrease between baseline and 6 months (Figure
701 7G).

702 For the comparison 6 months versus baseline, the results for LEfSe in the group who did not
703 outgrow their CMA showed large overlap with the top 10 LMM results ordered by unadjusted
704 p-values (Table S27). However for the other comparisons, the overlap was only moderate
705 (Table S27).

706 Finally, also human protein classes that changed between visits in both allergy groups were
707 found:plakins, proline-rich proteins, glycoside hydrolases and actin-binding proteins

(increase between baseline and 12 months, see Figures 7B and H); fatty acid binding proteins, carboxypeptidases, fused gene family and glyceraldehyde-erythrosephosphate dehydrogenase family (increase between 6 months and 12 months, see Figures 7C and I); filaggrins (increase between baseline and 12 months and between 6 months and 12 months, Figures 7B, C, H, I).

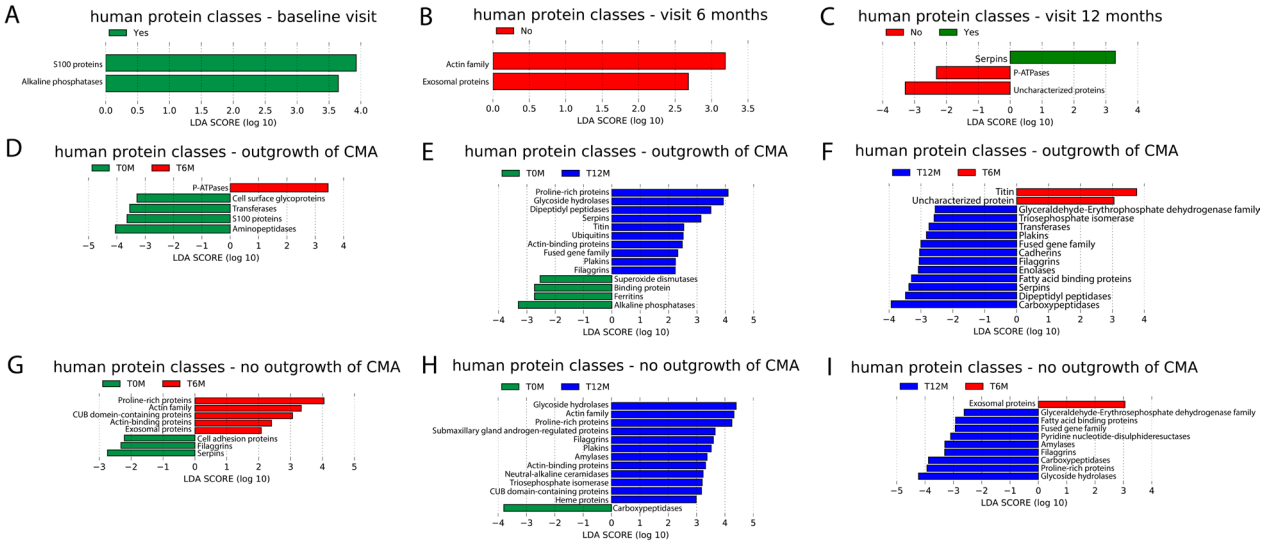

Figure 7. LEfSe analysis of human protein classes using an alpha value of 0.05 for the factorial Kruskal-Wallis test among classes and a threshold of 2.0 on the logarithmic LDA score for discriminative features. (A-C) Using outgrowth of CMA at visit 12 months as class, Yes: outgrowth of CMA at 12 months; no: no outgrowth of at 12 months. Plot of discriminative features at (A) baseline visit; (B) visit 6 months; (C) visit 12 months. (D-F) Pairwise comparison between visits within the group with outgrowth of CMA, using visit as class. (D) visit 6 months (T6M) versus baseline (T0M); (E) visit 12 months (T12M) versus baseline (T0M); (F) visit 12 months (T12M) versus visit 6 months (T6M). (G-I) Pairwise comparison between visits within the group with persistent CMA, using visit as class. (G) visit 6 months (T6M) versus baseline (T0M); (H) visit 12 months (T12M) versus baseline (T0M); (I) visit 12 months (T12M) versus visit 6 months (T6M).

Correlation between human and microbial proteins

To determine correlation between human and microbial proteins, Spearman correlations between human proteins of the top 10 protein classes and microbial proteins of the core taxa were calculated. Several significant correlations, of which the majority are weak, were found (Table S28). Significant moderate correlations (> 0.4) were observed between carboxypeptidases and *Coriobacteriaceae* (correlation =0.4598, p-value < 1e-04), and Transthyretin / hydroxyisourate hydrolases and *Coriobacteriaceae* (correlation =0.4603, p-value < 1e-04).

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

To determine the contribution of environmental and clinical factors to the variation in the gut microbiota, microbial proteome and human proteins, redundancy analysis (RDA) was performed. Outgrowth of CMA significantly explains variation in 16S rRNA gene based taxonomic profiles at visit 6 months (Table S29). However, compared to other significant features, the contribution of outgrowth of CMA is low (shorter edge in RDA plot, see Figure S6). Therefore, ordination plots coloured by outgrowth of CMA do not show a clear separation between the subjects with outgrowth of CMA and those with persistent CMA.
To determine the contribution of outgrowth of CMA to the variation in the gut microbiota and microbial proteome after adjusting for the other environmental variables, partial RDA was performed. The results show that the contribution is low and only significant for the 16S rRNA gene-based taxonomy (% variance explained by outgrowth of CMA = 6.37%, adjusted p-value = 0.043). The partial RDA plot shows that after removing the effect of other environmental factors, a more clear separation between the allergy classes (outgrowth vs persistent CMA) could be observed (Figure S7).
For the other visits, the microbial proteome profiles, the protein-based taxonomic profiles, the microbial protein functional profiles and the human protein profiles, the contribution of outgrowth of CMA to the variation in the data was not significant (Tables S29 and S30).

Contribution of human proteins on variation in the microbial proteome

To determine human proteins that significantly explain variation in microbial proteome and protein-based taxonomy profiles, another RDA analysis was performed. Table S31 shows that a considerable fraction of the variance in the microbial proteome and protein-based taxonomy profiles can be explained by human proteins at all visits. No clear grouping of the samples by allergy group was observed in any of the RDA plots (Figure S8).

Discussion

In this study, 16S rRNA gene amplicon sequencing and metaproteomics were applied on faecal samples in a cohort of infants with CMA at baseline visit, of which some showed outgrowth of CMA after 12 months and others did not. The aim was to get insight in outgrowth of CMA.
Apart from analysing the microbiome, environmental and clinical factors associated with outgrowth of CMA were determined. Our results show that among infants with that did not outgrow their CMA, a significantly higher proportion of infants with parental allergy was observed compared to infants that show outgrowth of CMA. This is consistent with previous

findings of Xinias et al., who showed that the risk of not acquiring outgrowth of CMA at 1 year of age significantly increased in infants with a family history of atopy [43]. Furthermore, a significantly lower SCORAD at baseline was observed in infants that show outgrowth of CMA compared to infants that do not. However, the SCORAD of the majority of the infants in this study is lower than 25, corresponding to mild severity of atopic dermatitis [44]. This suggests that in this cohort, CMA does not manifest mainly through atopic symptoms.

The faecal metaproteome was dominated by bacterial proteins from *Bifidobacteriaceae* and *Lachnospiraceae*. Proteins for these two families were also observed in a metaproteomics study of Kingkaw et al. [17] on atopic dermatitis, but the fraction of these taxa was lower than in our study. There are two possible explanations for the differences between our study and Kingkaw et al. First, the infants in our study have another CMA phenotype than atopic dermatitis. Second, the protein identification and quantification of Kingkaw et al. was based on a protein database derived from earlier studies of the gut microbiome which included only 10 bacterial families. In our study, we constructed an in-house proteomics database based on the 16S rRNA gene amplicon sequencing data consisting of 19 bacterial families. Third, the cohort in the study of Kingkaw et al. included both healthy and allergic children, while our study includes only children with CMA at baseline. The gut microbiome of infants with food allergy in their first year of life is known to be dominated by *Lachnospiraceae*, while this family is less abundant in the healthy gut microbiome [45].

Six common taxa were found in the top 10 of both 16S rRNA gene sequencing and metaproteomics: *Bacteroidaceae*, *Bifidobacteriaceae*, *Enterobacteriaceae*, *Erysipelatoclostridiaceae*, *Lachnospiraceae* and *Ruminococcaceae*. With the exception of *Erysipelatoclostridiaceae*, these are all common constituents of the healthy gut microbiome [46,47]. *Clostridiaceae*, *Coriobacteriaceae*, *Enterococcaceae* and *Tannerellaceae* were only among the top 10 taxa of protein-based microbial taxonomic profiles. The first two are common taxa in the gut microbiome [46]. The presence of increased levels of *Tannerellaceae* in the gut microbiome has been associated with the administration of synbiotics [48]. However, protein-based relative abundances of *Tannerellaceae* in the infants that received the synbiotic were lower than in the other infants (synbiotics: median = 0.003, interquartile range: 0.001-0.006; no synbiotics: median = 0.005, interquartile range: 0.003-0.011). This means that its appearance in the top 10 taxa in this study cannot be explained by the inclusion of 23 infants (59% of the infants) which received a formula with synbiotics. Taxa that were only among the top 10 in 16S rRNA gene-based profiles were *Akkermansiaceae*, *Erysipelotrichaceae*, *Streptococcaceae* and *Veillonellaceae*. The latter three are common constituents of the gut microbiome [46], while increased levels of *Akkermansiaceae* have been related with synbiotics interventions [48]. However, in this study, 16S rRNA gene-based relative abundances of *Akkermansiaceae* were lower in the infants that received the synbiotic than in those that did not (synbiotics: median = 0.00019, interquartile range: 0.00005-0.05522; no synbiotics: median = 0.00266, interquartile range: 0.00016-0.60490), so the occurrence of this family in the top 10 is not due to the synbiotic treatment of 23 infants. Differences between 16S rRNA gene-based and protein-based microbial composition have been observed in a previous study [49], for which two possible explanations were given

814 by the authors. First, it could be that the 16S rRNA gene-based microbial composition is not
815 representative for microbial protein abundance. Second, we measure a lot less proteins than
816 DNA sequences. Third, protein expression of bacteria can be different. Some bacteria will be
817 more active in producing proteins. Forth, in the pool of proteins that we extracted, we are
818 also depending on the stability of the proteins.
819 Linear mixed model (LMM) analysis of core taxa and LEfSe analysis on the whole data were
820 applied on the 16S rRNA gene sequencing and metaproteomics data to compare infants who
821 outgrew their CMA with infants that did not.
822 Our study showed significantly higher 16S rRNA gene-based relative abundances of
823 *Lachnospiraceae* at baseline in infants that did not outgrow their CMA compared to infants
824 who outgrew their CMA. In contrast, a previous study reported higher levels of *Lachnospira*
825 *pectinoschiza* in infants that outgrew their allergy [50]. However, this study focuses on children
826 with a mean age of 57.8 months at baseline, while the infants in our study were 15–25
827 months old at visit 12 months (3-13 months old at baseline). Moreover, higher relative
828 abundance of a family does not mean that the relative abundance of each specific species
829 belonging to that family had to be higher. Furthermore, two families that, to the best of our
830 knowledge, have not been related to CMA (*Carnobacteriaceae* and *Peptostreptococcales-*
831 *Tissierellales_fa*) showed higher relative abundances at baseline in the group that did not
832 outgrew their CMA.
833 In infants who outgrew their CMA, higher relative abundances of *Enterococcaceae and*
834 *Leuconostocaceae* at baseline visit were observed. Higher levels of the genus *Enterococcus*
835 have been reported in non-allergic infants when compared to allergic infants [51]. In our study,
836 the genus Enterococcus has higher relative abundances at baseline in the group that
837 outgrew their CMA compared to the other group (outgrowth of CMA: median = 0.00142,
838 IQR: 0.00029-0.00544; no outgrowth of CMA: median = 0.00006, IQR: 0.00000-0.00310).
839 The higher relative abundances of *Aerococcaceae* and *Pasteurellaceae* at visit 6 months in
840 the group who did not outgrew their CMA are in contradiction with Kourosh et al.[52], who
841 observed lower Pasteurellales in the allergic group compared to the control. However, this
842 study was not restricted to early life but ranged from birth to the age of 18 years old.
843 Infants who outgrew their CMA showed higher relative abundances of *Eubacteriaceae* at visit
844 6 months than those who did not. Bunyavanich et al.[53] reported an association between milk
845 allergy resolution and higher levels of the genus *Eubacterium*. In our study, higher relative
846 abundances of *Eubacterium* were also found at visit 6 months in the infants who outgrew
847 their CMA compared to the other group of infants (outgrowth of CMA: median = 0.00083,
848 IQR: 0.00000-0.00229; no outgrowth of CMA: median = 0.00000, IQR: 0.00000-0.00027).
849 At visit 12 months, higher *Bacteroidaceae* were observed in infants who did not outgrew
850 their CMA. Higher levels of the phylum Bacteroidetes were reported earlier in infants with
851 persistent CMA [53]. For the group who outgrew their of CMA, higher relative abundances were
852 found at visit 12 months for three families that, to the best of our knowledge, have not been
853 associated earlier with outgrowth of CMA: *Rikenellaceae*, *Actinomycetaceae* and
854 *Eggerthellaceae*.

When comparing visits within allergy groups, we found higher *Ruminococcaceae* at 12 months compared to baseline in the group with outgrowth of CMA, while no significant changes in this family were observed in infants with persistent CMA. Bunyavanich et al. [53] reported increased levels of the genus *Ruminococcus* in infants with resolved CMA compared with persistent CMA. In our study, the median relative abundance increased from 0.00000 to 0.00036 between baseline and 12 months in the group who outgrew their CMA, while in the other group the medians at baseline and 12 months were equal (both equal to 0.00003). Furthermore, our analyses identified higher *Bacteroidaceae* at 12 months compared to baseline in the group who did not outgrow their CMA, while no significant changes were found in the group who outgrew their CMA. Higher levels of Bacteroidetes were associated earlier with persistent CMA [53].

In both allergy groups, lower *Enterobacteriaceae* were observed at 12 months compared to baseline visit. As *Enterobacteriaceae* are generally more abundant in young infants and decrease with age[2], this finding could be the result of that. Decreased levels of *Enterobacteriaceae* have been previously reported in infants with CMA that received an amino acid-based formula during a period of 6 months [15].

In both allergy groups, the results showed higher relative abundances of the families *Butyricicoccaceae* and *Monoglobaceae* at visit 12 months compared to baseline, which suggests that these families change when infants growing older.

In our study, relative abundances of *Eubacteriaceae*, *Oscillospiraceae* and *Tannerellaceae* increased between baseline and 6 months in the group who outgrew their CMA, but not in the other group. To the best of our knowledge, these three families have not been associated with milk allergy resolution earlier.

Furthermore, we found lower *Enterococcaceae*, *Staphylococcaceae* and *Saccharimonadeceae* at visit 6 months compared to baseline in the group who outgrew their CMA, but not in the other group. Significantly lower *Enterococcaceae* were reported earlier in healthy infants compared to infants with food sensitization [54]. A systematic review [55] reported contradictory results in the literature for *Staphylococcaceae* in relation to food sensitization (i.e. some studies relate food sensitization with higher *Staphylococcaceae*, while others found lower *Staphylococcaceae*).

Several families only increase between baseline and 12 months in the group who outgrew their CMA, but not in the other group: *Christensenellaceae*, *Oscillospiraceae*, *Oscillospirales_fa*, *Clostridia UCG 014* family, *Acidaminococcaceae*, *Rikenellaceae* and *Lachnospiraceae*. Higher relative abundances of *Rikenellaceae* were also reported in healthy infants compared to infants with food sensitization by Chen et al. [54]. The same study also reported higher levels of *Acidaminococcaceae*. However, the results were not significant.

A decrease in *Micrococcaceae* between baseline and visit 12 months was found in the group who outgrew their CMA, but not in the other group. To the best of our knowledge, this family has not been previously related to CMA resolution.

In infants who outgrew their CMA, but not in the other group, higher relative abundances of *Rikenellaceae*, *Actinomycetaceae* and *Prevotellaceae* at visit 12 months compared to visit 6 months were identified. Higher *Prevotellaceae* were related earlier to CMA resolution [53].

897 Remarkably, when comparing visit 12 months to baseline, higher *Prevotellaceae* were found
898 in the group who did not outgrew their CMA, which contradicts with previous findings [53].
899 In the group who did not outgrew their CMA, but not in the other group, higher
900 *Leuconostocaceae* at 12 months compared to baseline, and higher *Selemonadaceae* at 12
901 months compared to 6 months were observed. As far as we are aware, these two families
902 have not been related to persistent CMA earlier.
903 A much lower amount of significant results were identified by the metaproteomics analysis at
904 the taxonomy level compared to the analysis of the 16S rRNA gene-based taxonomy. This
905 suggests that changes related to outgrowth of allergy mainly occur at the level of the 16S
906 rRNA gene-based taxonomic classification and less at the protein-based taxonomic
907 classification level. When comparing allergy groups within visits, the only significant result at
908 the family level were higher *Eggerthellaceae* proteins at baseline and higher *Veillonellaceae*
909 proteins at visit 6 months in infants who did not outgrow their CMA compared to the group
910 who outgrew their CMA. To the best of our knowledge, these results have not been reported
911 earlier.
912 Similar as for 16S rRNA gene sequencing, the results for metaproteomics showed an
913 decrease in relative abundance of *Enterobacteriaceae* proteins between baseline and 12
914 months in both allergy groups.
915 The results of the metaproteomics analysis reveal microbial metabolic processes which are
916 expected to play a crucial role in the mechanisms underlying outgrowth of CMA. For most of
917 the processes, multiple bacterial families contribute to their increase over visits in the group
918 who outgrew their CMA (Table S32). In infants who outgrew their CMA, increased relative
919 abundances of *Rikenellaceae* and *Coriobacteriaceae* proteins between baseline and 12
920 months can be related to an increase in relative abundance of proteins involved in pyruvate
921 metabolism and fructose and mannose metabolism respectively (Table S32).
922 In general, the analysis on microbial protein functional classes did identify fewer significant
923 differences between allergy groups and visits compared to the analysis of the 16S rRNA
924 gene-based taxonomy, suggesting that changes related to outgrowth of CMA occur at the
925 16S rRNA gene-based taxonomic level and less at the protein functional level.
926 In this study, we also determined differences in human protein classes between allergy
927 groups and visits. For children who did not outgrew their at 12 months, significantly lower
928 levels of human immunoglobulins were observed at visit 12 months compared to the
929 baseline visit. Furthermore, higher relative abundances of proline-rich proteins at 12 months
930 compared to baseline, and lower relative abundances of S100 proteins at 6 months
931 compared to baseline were observed in the group who outgrew their allergy, but not in the
932 other group.
933 The results of the RDA analysis confirm that there is a small contribution of the microbiome
934 to outgrowth of CMA, and that this contribution mainly occurs at the 16S rRNA gene-based
935 taxonomy level.
936 Our study has several limitations. First, the study has a wide age range at baseline visit (3-
937 13 months). Given the large changes that the infant gut microbiome undergoes between 3
938 and 13 months, among others related to weaning, this wide range could affect the results.

939 Large differences in baseline microbiota and in its development over time are expected
940 between infants of 3 and 13 months old. Moreover, infants of similar age may be included at
941 baseline and 6 months, and at 6 and 12 months. Age-related variability within visits may
942 have confounded the comparisons between visits, potentially obscuring changes related to
943 age and outgrowth of CMA. Overall, adding age as fixed effect to the LMM resulted in worse
944 goodness of fit (higher AIC) and a lower number of significant differences, in particular
945 between visits. Second, the size of the cohort is small, and as a consequence the results
946 have to be interpreted with caution. Validating the results in a similar but larger cohort is a
947 direction for future research. Third, the small size of the cohort made it difficult to detect
948 significant features from a complete list of taxa or protein classes using methods correcting
949 for multiple testing, like LMM analysis. Methods based on effect size rather than statistical
950 significance after multiple testing correction, like LEfSe, are more sensitive to detect
951 discriminative features in a small cohort. Their drawback is the risk for detecting false
952 positives. Therefore, we conducted a LMM analysis on the complete list of families (in
953 addition to the analysis of the core taxa) and the complete lists of microbial functional and
954 human protein classes (in addition of the analysis of the top 10). As expected, the number
955 of significant features found by this LMM analysis was low, but we noticed that the majority
956 of the results found by LEfSe were in the top ranked list of LMM-based unadjusted p-values.
957 This indicates concordance between the LMM and LEfSe results, and suggests that the
958 majority of differential features obtained by LEfSe could have also been detected by LMM in
959 case of availability of a larger cohort.
960 In summary, we can conclude microbiome differences related to outgrowth of CMA can be
961 mainly identified at the level of the 16S rRNA gene taxonomic level, and to a lesser extent at
962 the protein-based microbial taxonomy and functional level. The overall contribution of the
963 microbiome to outgrowth of CMA is low.
964

965 # Data availability

966 Clinical data are available from Danone Nutricia research upon reasonable request (contact:
967 Harm Wopereis, Danone Nutricia Research, Utrecht, The Netherlands,
968 Harm.Wopereis@danone.com).
969 The 16S rRNA-gene sequencing for this study have been deposited in the European
970 Nucleotide Archive (ENA) at EMBL-EBI under accession number PRJEB56782
971 (https://www.ebi.ac.uk/ena/browser/view/PRJEB56782).
972 The mass spectrometry proteomics data have been deposited to the ProteomeXchange
973 Consortium via the PRIDE [56] partner repository with the dataset identifier PXD037190.
974 The R code for the statistical calculations made in this study has been deposited to a public
975 repository on GitLab (https://git.wur.nl/afsg-microbiology/publication-supplementary-
976 materials/2022-hendrickx-et-al-earlyfit_presto_allergy_study).
977
978
979

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 Mischke, M. *et al.* Specific synbiotics in early life protect against diet-induced obesity in adult mice. *Diabetes Obes Metab* **20**, 1408-1418, doi:10.1111/dom.13240 (2018).

20 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).

21 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).

22 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).

23 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).

24 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).

25 Amir, A. *et al.* Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. *Msystems* **2**, doi:ARTN e00191-16 10.1128/mSystems.00191-16 (2017).

26 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).

27 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).

28 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).
29 Johnson, M. *et al.* NCBI BLAST: a better web interface. *Nucleic Acids Res* **36**, W5-9, doi:10.1093/nar/gkn201 (2008).
30 UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res* **49**, D480-D489, doi:10.1093/nar/gkaa1100 (2021).
31 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).
32 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).
33 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).
34 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).
35 R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing, Vienna, Austria, 2019).
36 Lahti, L. & Shetty, S. microbiome R package. (2012-2019).
37 Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research* **28**, 27-30 (2000).
38 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).
39 Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J. A. & Pawlowsky-Glahn, V. Logratio analysis and compositional distance. *Mathematical Geology* **32**, 271-275 (2000).
40 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).
41 Oksanen, J. *et al.* (2020).
42 Segata, N. *et al.* Metagenomic biomarker discovery and explanation. *Genome biology* **12**, 1-18 (2011).
43 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).
44 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).
45 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).
46 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).
47 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).
48 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).
49 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).
50 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).
51 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).
52 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).
53 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).
54 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).
55 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).
56 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.

Additional information

Supplementary material for this paper is available at XXX.

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

1142 **Knipping[3], Johan Garssen[3,11], Valerie Trendelenburg[12], Robert Pesek[13], Carla M.**
1143 **Davis[14], Antonella Muraro[15], Mich Erlewyn-Lajeunesse[16], Adam T. Fox[17], Louise J.**
1144 **Michaelis[18], Kirsten Beyer[19], Lee Noimark[20], Gary Stiefel[21], Uwe Schauer[22], Eckard**
1145 **Hamelmann[22], Diego Peroni[23], and Attilio Boner[23]**

1146 [4] Pediatric Allergy and Clinical Immunology Research Unit, Division of Allergy and
1147 Immunology, Department of Pediatrics, Faculty of Medicine, Chulalongkorn University, King
1148 Chulalongkorn Memorial Hospital, Thai Red Cross Society, Bangkok, Thailand
1149 [5] Allergy and Immunology, Department of Pediatrics, New York University Langone Health,
1150 New York, US
1151 [6] Department of Pediatrics, Gastroenterology and Nutrition, Collegium Medicum, University of
1152 Warmia and Mazury, Olsztyn, Poland
1153 [7]St Marien Hospital,Bonn, Germany
1154 [8]Pediatric Allergy and Immunology Division, Department of Pediatrics, Faculty of Medicine
1155 Ramathibodi Hospital, Mahidol University, Bangkok, Thailand
1156 [9]Allergy Service, Department of Paediatric Medicine, KK Women's & Children's Hospital,
1157 Singapore
1158 [10]Department of Pediatrics, Faculty of Medicine, Prince of Songkla University, Hat Yai,
1159 Thailand
1160 [11]Utrecht Institute for Pharmaceutical Sciences, Utrecht University, Utrecht, The Netherlands
1161 [12] Department of Pediatric Pneumology, Immunology, and Intensive Care Medicine, Charité
1162 Universitätsmedizin Berlin, Berlin, Germany
1163 [13]Arkansas Children's Hospital, Little Rock, US
1164 [14]Texas Children's Hospital, Baylor College of Medicine, Houston, US
1165 [15]Food Allergy Referral Centre, Padua University Hospital, Padua, Italy
1166 [16]University Hospitals Southampton, Southampton, United Kingdom
1167 [17]Guy's and St Thomas' NHS Foundation Trust, London, United Kingdom
1168 [18]Great North Children's Hospital, Newcastle Upon Tyne Hospitals NHS Foundation Trust,
1169 Newcastle Upon Tyne, United Kingdom
1170 [19]Department of Pediatric Pneumology, Immunology, and Intensive Care Medicine, Charité
1171 Universitätsmedizin Berlin, Berlin, Germany
1172 [20]Barts/Royal London Hospital, London, England, United Kingdom
1173 [21]Leicester Royal Infirmary, Leicester, England, United Kingdom
1174 [22]Ruhr-Universitat Bochem im St Josef-Hospital, Bochum, Germany
1175 [23]University Hospital Verona, Verona, Italy