Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Early-life Gut Virome Maturation and Association with Celiac Disease Development: a Nested Case-Control Study

Ondrej Cinek

Ondrej.Cinek@lfmotol.cuni.cz

Charles University in Prague, University Hospital Motol and Homolka https://orcid.org/0000-0002-0526-8714

Ulrika Boulund

COPSAC, Copenhagen Prospective Studies on Asthma in Childhood, Copenhagen University Hospital - Herlev and Gentofte, Copenhagen, Denmark

Katerina Chudá

Charles University, Second Faculty of Medicine, Department of Medical Microbiology

Shiraz Shah

Copenhagen University Hospital Herlev-Gentofte https://orcid.org/0000-0002-4665-577X

German Tapia

Department of Chronic Diseases, Norwegian Institute of Public Health, Oslo

Klara Hubackova

Second Faculty of Medicine, Charles University, and University Hospital Motol and Homolka

Michaela Čermáková

Charles University, Second Faculty of Medicine, Department of Medical Microbiology

Jutta Laiho

Tampere University https://orcid.org/0000-0002-5443-4651

Jussi Lehtonen

Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland

Kalle Kurppa

Tampere University https://orcid.org/0000-0003-4757-2164

Katri Lindfors

Tampere University

Mikael Knip

University of Helsinki https://orcid.org/0000-0003-0474-0033

Jorma Toppari

Turku University Hospital https://orcid.org/0000-0003-2228-334X

Riitta Veijola

University of Oulu https://orcid.org/0000-0002-6557-270X

Klaus Bønnelykke

University of Copenhagen https://orcid.org/0000-0003-2003-1018

Ketil Størdal

Department of Paediatric Research, Institute of Clinical Medicine, Faculty of Medicine, University of Oslo, Oslo, Norway

Heikki Hyöty

Tampere University https://orcid.org/0000-0003-0370-4145

Jakob Stokholm

Copenhagen Prospective Studies on Asthma in Childhood https://orcid.org/0000-0003-4989-9769

Lars Stene

Norwegian Institute of Public Health https://orcid.org/0000-0002-8434-119X

Jonathan Thorsen

COPSAC https://orcid.org/0000-0003-0200-0461

Article

Keywords: microbiome, infant, cohort, bacteriophages, autoimmunity

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11713360/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

Early-life Gut Virome Maturation and Association with Celiac Disease Development: a Nested Case-Control Study

Ulrika Boulund*[1,2], Katerina Chuda*[3], Shiraz A. Shah[1], German Tapia[4], Klara Hubackova[3], Michaela Cermakova[3], Jutta E. Laiho[5], Jussi P. Lehtonen[5], Kalle Kurppa[5,6], Katri Lindfors[5], Mikael Knip[7,8], Jorma Toppari[9], Riitta Veijola[10], the HEDIMED Investigator group, Klaus Bønnelykke[1,11], Ketil Størdal[12], Heikki Hyöty[5,8,13], Jakob Stokholm[1,14], Lars C. Stene[4], Ondrej Cinek**[3] , Jonathan Thorsen**[1,11]

1. COPSAC, Copenhagen Prospective Studies on Asthma in Childhood, Copenhagen University Hospital - Herlev and Gentofte, Copenhagen, Denmark
2. Faculty of Health and Medical Sciences, Department of Public Health, University of Copenhagen, Copenhagen, Denmark
3. Department of Medical Microbiology, Second Faculty of Medicine, Charles University; University Hospital Motol and Homolka, Czechia
4. Department of Chronic Diseases, Norwegian Institute of Public Health, Norway
5. Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland
6. University Consortium of Seinäjoki, Seinäjoki, Finland
7. Research Program for Clinical and Molecular Medicine, University of Helsinki, Helsinki, Finland
8. Department of Pediatrics, Tampere University Hospital, Tampere, Finland.
9. Research Centre for Integrative Physiology and Pharmacology; Centre for Population Health Research, InFLAMES Research Flagship Centre, Institute of Biomedicine, University of Turku, Turku, Finland; Department of Pediatrics, Turku University Hospital, Turku, Finland
10. Department of Pediatrics, Research Unit of Clinical Medicine, Medical Research Center, University of Oulu, Oulu, Finland; Department of Children and Adolescents, Medical Research Center, Oulu University Hospital, Oulu, Finland
11. Department of Clinical Medicine, University of Copenhagen, Denmark
12. Department of Paediatric Research, Institute of Clinical Medicine, Faculty of Medicine, University of Oslo, Oslo, Norway
13. Fimlab Laboratories, Tampere, Finland
14. Department of Food Science, University of Copenhagen, Copenhagen, Denmark

*Shared first authors; **Shared senior authors

Key words: microbiome, infant, cohort, bacteriophages, autoimmunity

Correspondence to:
Prof. Ondrej Cinek, Ph.D.
Department of Medical Microbiology
University Hospital Motol and Homolka, V Uvalu 84, Prague; CZ-150 06, Czechia
ondrej.cinek@Lfmotol.cuni.cz
Phone +420 2 2443 2026

Abstract

The infant gut virome development and its role in autoimmune diseases are largely unexplored. Celiac disease (CeD) is increasing in incidence worldwide, and we recently reported that a higher frequency of enterovirus or parechovirus gut infections precedes CeD autoimmunity. In this nested case-control study of two independent prospective birth cohorts we analysed the overall virome composition before the onset of CeD. In total, 2,094 stool samples were collected monthly between 3-36 months of age from 41 children who were subsequently diagnosed with CeD and 76 matched controls. These samples were subject to viral-like particle enrichment and shotgun metagenomic sequencing of the DNA virome. The overall virome composition was not associated with CeD, but a multivariate model based on the abundance of viral families with a virulent lifestyle could modestly predict CeD status. The prediction scores of this model were significantly associated with CeD in a non-linear manner, suggesting that bacteriophages with a virulent lifestyle are involved in the risk of developing CeD.

Introduction

The human gut microbiome develops rapidly in the first years of life[1]. The viral component of the gut microbiome (virome) is composed predominantly of bacteria-infecting viruses (bacteriophages, phages) that play a pivotal role in shaping the bacterial community structure[2]. Virulent phages infect and lyse host bacteria, while temperate phages integrate into the bacterial chromosome, only lysing their host bacterium when certain conditions are met. The virome can play an independent role in the maturation of the immune system[3,4], and has recently been implicated in childhood disease[5,6]. There are also human-infecting viruses in the gut, some cause disease, whereas others are speculated to be part of the commensal microbiome[7,8,9]. Yet, our understanding of the development of the overall virome composition in early life and its role in health is still limited[1].

Celiac disease (CeD) is a multifactorial autoimmune disorder characterized by a complex interplay between genetic susceptibility and environmental factors, specifically the ingestion of gluten. CeD is increasing in incidence globally[10–12]. It is associated with the HLA-DQ2 or HLA-DQ8 molecules, which are present in more than 95% of CeD patients[10–12]. Approximately 40-50% of the North European population are carriers of either HLA-DQ2, DQ8, or both. As the population-level risk of developing CeD is 1-3%, the majority of the carriers never develop the disease[12–16]. This highlights a critical gap in our understanding of the disease triggers. We and others recently reported that a higher frequency of enterovirus[17,18] and parechovirus[7] gut infections precedes seroconversion to positivity for tissue transglutaminase autoantibodies (tTGA), a reliable preclinical serological marker of celiac disease autoimmunity (CDA) that precedes the overt disease. Other human gut viruses frequently occurring in infancy were also investigated, showing no association[2,7,19]. The overall composition of the gut virome in the context of CeD development has not been investigated before.

Almost two decades ago, the first metagenome-sequenced virome of a single healthy one-week-old infant was described[2]. The virome consisted mainly of phages, was of low diversity, and exhibited dramatic changes in the first two weeks of life. Since then, a few studies describing the gut virome development have been published[20–26], demonstrating a large shift in the overall virome composition in early-life[27], and a decrease in the inter-individual variability with age[22]. The original colonizing phages are likely induced prophages from pioneering bacteria[28], exhibited by a high proportion of temperate phages in the earliest samples[22,27]. Previous literature suggests that the rapid development of the gut virome and bacteriome in early-life closely follows the maturation of the immune system[29]. A major challenge in analysing viromics data has been the so-called viral dark matter, though recent

advancements in bioinformatics now allow for greater insight into previously unknown gut viruses. We previously identified 19 unknown viral families that were associated with a protective effect against the development of childhood asthma[5,30].

The aim of this study is to improve the understanding of the early-life gut virome development and its role in the development of CeD. Here, we leverage two independent prospective birth cohorts with fecal samples collected before the onset of clinical CeD, to investigate the maturation of the virome in children who later develop CeD compared to matched controls.

Results

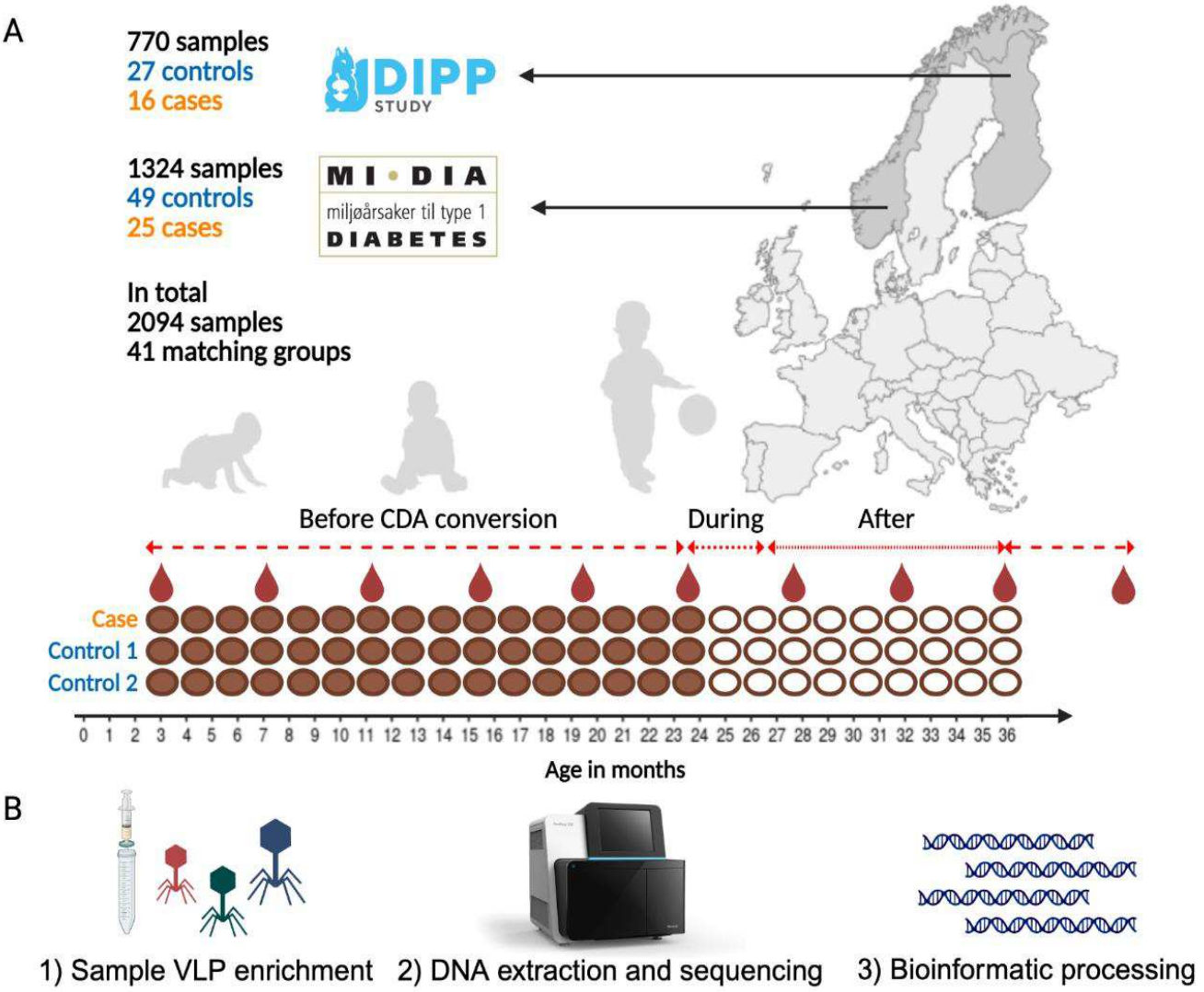
We analyzed the DNA gut virome in early life with 2,094 fecal samples from 117 children, through the ages of 3 months to 3 years (**Figure 1**). The study was designed as a nested case control study in two prospective birth cohort studies; DIPP (Type 1 Diabetes Prediction and Prevention study, Finland) and MIDIA (Environmental causes of Type 1 Diabetes, Norway) of children with increased HLA-associated risk of Type 1 Diabetes and CeD[31,32]. Blood samples were collected approximately every three months in the first year, and 1-2 times annually thereafter (see methods for details). The CDA was identified by testing the blood samples for tTGA, and CeD was diagnosed by ESPGHAN criteria (see methods for details). Cases (N=41) were matched with CDA-negative controls (N=76) by date and place of birth (**Table 1, Suppl. Figure 1, Suppl. Table 1**). Stool samples were collected approximately every month (see methods for details) and underwent viral-like particle enrichment and metagenomic sequencing. We agglomerated viral operational taxonomic units (vOTUs) to viral family-level clades (VFCs) and to predicted bacterial host genus.

Figure 1. Study design.

A) A nested case-control study was established in the Finnish DIPP and Norwegian MIDIA prospective birth cohorts. Children with increased HLA-associated risk of celiac disease were followed with monthly fecal samples from 3 to 36 months of age (brown and empty circles), and blood samples every three months in the first 12-18 months of life and annually thereafter (see methods for details). Children developing celiac disease were matched with controls by date and place of birth and duration of follow-up. In this study, only the fecal samples collected before seroconversion were used (filled brown circles).

B) Viral-like particles from fecal samples were enriched by filtering and centrifugal separation, and the DNA underwent shotgun metagenomic sequencing after unbiased amplification by sequence-independent, single-primer amplification (SISPA).

87


A
770 samples
27 controls
16 cases
DIPP STUDY
1324 samples
49 controls
25 cases
MI • DIA
miljøårsaker til type 1
DIABETES
In total
2094 samples
41 matching groups
Before CDA conversion
During
After
Case
Control 1
Control 2
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36
Age in months
B
1) Sample VLP enrichment
2) DNA extraction and sequencing
3) Bioinformatic processing

88

Table 1. Cohort characteristics.

	CeD Case	Control	P-value
Number of participants	41 MIDIA cohort: 25 (61%) DIPP cohort: 16 (39%)	76 MIDIA cohort: 49 (64.5%) DIPP cohort: 27 (35.5%)	0.60
Female, n (%)	20 (48.8)	34 (44.7)	0.57
Age of last negative tTGA (months, mean (SD))	33.6 (21.5)		
Born by C-section, n (%)	8 (19.5)	13 (17.1)	0.65
Have older siblings, n (%)	27 (65.9)	52 (68.4)	0.65
Number of gastrointestinal infections per month (mean (SD))	0.2 (0.2)	0.2 (0.2)	0.95
Number of antibiotics courses per month (mean (SD))	0.1 (0.1)	0.2 (0.3)	0.17
Age of introduction of gluten containing cereals (months, mean (SD))	6.2 (1.3)	6.0 (1.5)	0.33
Type 1 diabetes, n (%)	5 (12.2)	1 (1.3)	0.04
Birth year (median [IQR])	2005 [2002, 2006]	2005 [2002, 2006]	0.65

Birth season (number of participants) (mean (SD))			0.99
Autumn (Sep, Oct, Nov)	13 (31.7)	23 (30.3)	
Spring (March, April, May)	11 (26.8)	19 (25.0)	
Summer (June, July, Aug)	10 (24.4)	20 (26.3)	
Winter (Dec, Jan, Feb)	7 (17.1)	14 (18.4)	
Number of fecal samples per child (mean (SD))	17.8 (8.6)	17.9 (9.1)	0.66

All values refer to the samples collected before seroconversion for tTGA in the cases matched with controls. Gastrointestinal infections are based on questionnaires on symptomatic infections. SD = Standard deviation. IQR = interquartile range. tTGA = tissue transglutaminase autoantibodies, HLA = human leukocyte antigen. The Chi-squared test was used to test association between cases and controls for the categorical variables, a one-way analysis of variance (ANOVA) test was used for the continuous variables, though the Wilcoxon rank-sum test was used for birth year and birth month.

Composition and diversity of the infant gut virome

From the 2,094 successfully sequenced viromes, we derived 1.1 TB sequence data, averaging 3.4M +/- 0.6M reads per sample. These were assembled separately per sample and pooled into 1.6M and 1.8M contigs for MIDIA and DIPP respectively (**Suppl. Figure 2**). After clustering contigs at 95% average nucleotide identity (ANI), 380,000 contig representatives remained from either cohort. After running geNomad[33] on these, 9,800 and 4,700 contig representatives were flagged as viral from MIDIA and DIPP cohorts, respectively. These were then pooled with 11,660 taxonomically resolved vOTUs from the $COPSAC_{2010}$ infant gut viromes[30] and clustered at the viral species level (95% ANI) once more, followed by a final quality control. This procedure yielded 15,114 distinct vOTUs, of which 93% were bacteriophages. All vOTUs were then clustered *de novo* at the deeper viral family level using remote protein homology, yielding 511 VFCs. Of these, 231 VFCs corresponded to 183 viral families already defined as part of the $COPSAC_{2010}$ infant gut virome study published previously[30]. Of the 511 VFCs, 147 named and 53 unnamed VFCs had a prevalence >30% across all samples and were kept for further analysis (**Figure 2** and **Suppl. Figure 3**). All the top 30 largest viral families had been observed and named previously, now accounting for more than half of the total viral species-level diversity (**Figure 2C**). In contrast, most of the novel (yet unnamed) VFCs were smaller, rarer or less abundant. The temperate vs. virulent lifestyle of each vOTU was predicted using a machine-learning algorithm (**Suppl. Figure 2**), and the weighted average of this prediction was used to determine the VFC-level lifestyle. Temperate VFCs far outnumbered virulent VFCs in the data, although virulent VFCs were generally more abundant (**Figure 2A**). To capitalize on this biology, the VFCs were divided into viral subsets for the subsequent analyses throughout the manuscript; 1) all VFCs, 2) temperate VFCs or 3) virulent VFCs.

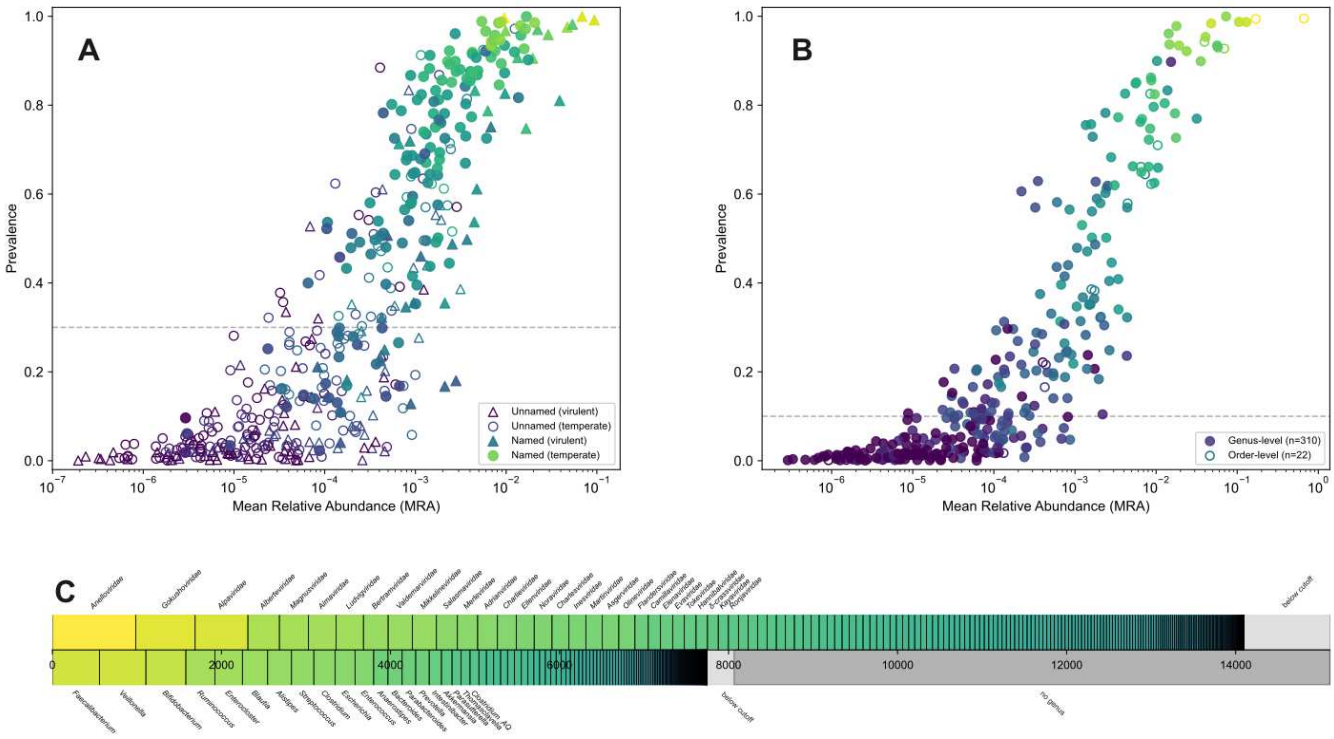
Bacterial host was bioinformatically predicted[34] for each vOTU. Multiple host predictions per vOTU were resolved by assigning the last common bacterial ancestor as the predicted host. A little over half of the vOTUs had specific host assignments at the bacterial genus-level (**Figure 2C**). Most of the remaining vOTUs had deeper bacterial rank assigned as hosts, e.g. at the bacterial order level, possibly reflecting their broader host ranges. The most frequent genus-specific hosts were *Faecalibacterium*, *Veillonella* and *Bifidobacterium*, with Bacteroidales being the most frequent host order (**Figure 2B** and **C**). The subset of bacterial host genera that were included for downstream analyses were required to have a sample prevalence of at least 10% in order to ensure adequate dynamic range and limit noise for model fitting.

Figure 2. Distribution of viral families and bacterial host genera in the cohort samples.

A) Mean relative abundance (MRA) versus prevalence for each viral family (VFC). Triangles denote virulent families, circles denote temperate families. Solid circles represent named infant gut viral families[30]. Open circles represent yet unnamed VFCs. Point color encodes species richness per VFC. The dashed line indicates the 30% prevalence cutoff used to filter the data set before association against child-level outcomes.

B) MRA versus prevalence of vOTUs when aggregated by their bacterial host genus. Only half of the vOTUs had host assignments at the bacterial genus-level. Solid circles represent bacterial genera, while open circles represent deeper bacterial orders, for vOTUs lacking a genus-level host assignment. The dashed line indicates the 10% prevalence applied before downstream association analyses

C) Stacked bar charts showing the composition of all 15,114 vOTUs by viral family (above axis) and bacterial host genus (below axis). Segments are colored and ordered by descending viral species-level richness; the gray segment at the right end of each bar collects families or genera below the applied prevalence filters. For the lower bacterial half, the darker gray "no genus" segment represents vOTUs that were excluded from further analysis because of a lacking host assignment at the bacterial genus level[30]. An interactive version of the figure with expandable families can be accessed online, for browsing the gene contents and downloading the genome of each virus: http://hedimed.crispr.dk.



The development of the gut virome in early life

The richness of the viromes (observed vOTUs) increased dramatically with age, especially from 3-12 months (**Suppl. Figure 4A,B)**. In the youngest samples (<4 months of age), the observed number of vOTUs was on average 451, ranging between [144-915], around 6 months of age 577, [136-1254], at 12 months 596 [339-1620], and at 24 months 1125 [270-2040]. This pattern was also mirrored in the observed VFCs (**Figure 3A, B**). Similarly, there was a distinct age-associated gradient of the overall virome family composition (beta diversity, Aitchison distance of the VFCs), suggesting that compositional differences were less pronounced after approximately 18-30 months (**Figure 3C**). This was tested more formally by assessing the Aitchison distance between consecutive samples within a single child, which revealed that consecutive samples in children above the age of 18 months had a significantly lower distance compared with earlier sampling intervals ($\beta = -0.41$, 95% confidence interval (CI) [−0.51;−0.31], $p<2x10^{-16}$). These observations were confirmed in stratified analyses across the two independent cohorts (**Suppl. Figure 4C,D**), confirming that this virome maturation is universal and not due to methodological or other cohort-specific factors. When analysing consecutive samples within a single child, it was also clear that a longer time interval between samples was associated with a larger compositional dissimilarity (**Suppl. Figure 7D**, linear mixed model (LMM) testing the Aitchison distance of consecutive samples of children).

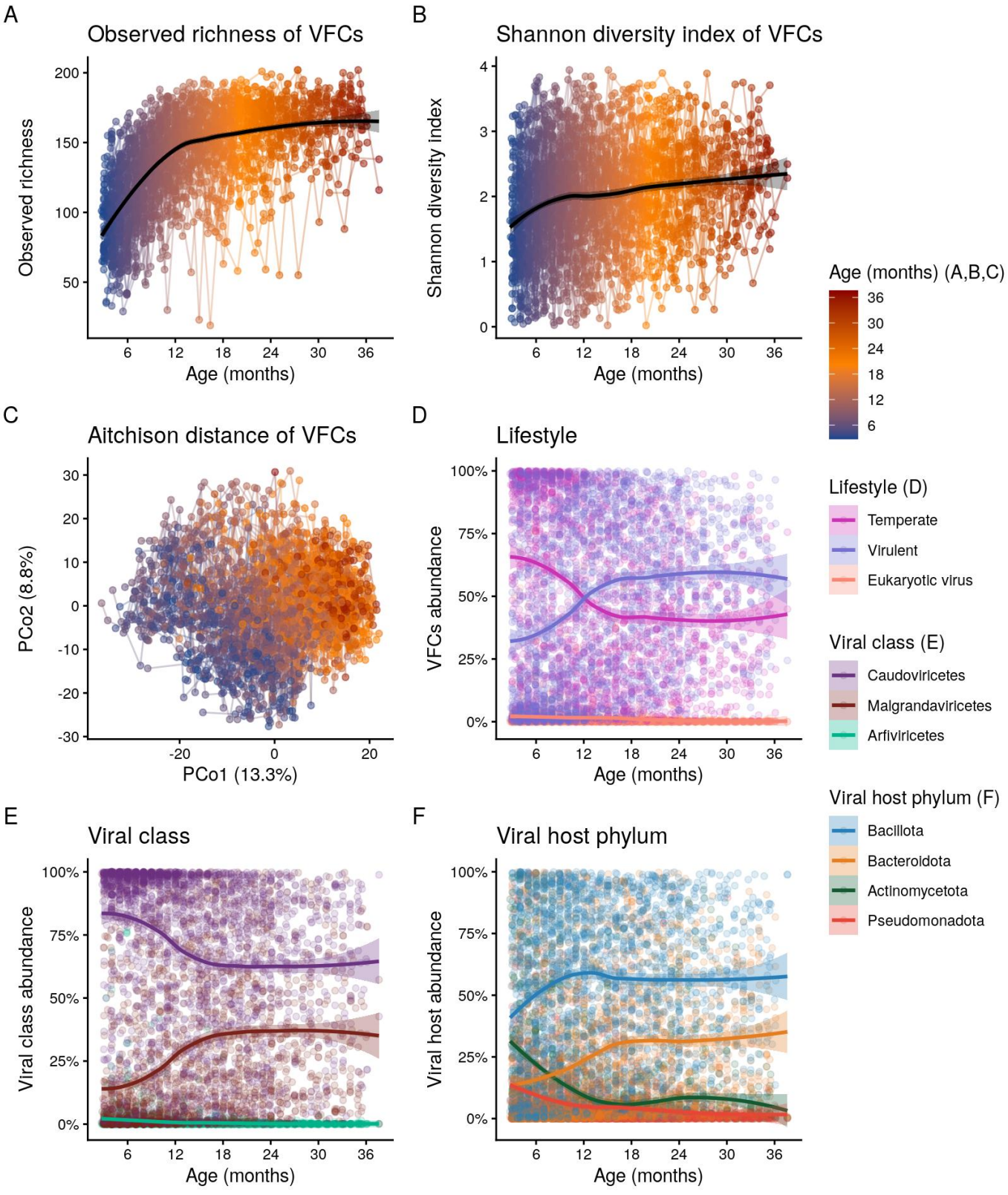
Early samples were dominated by temperate phage families (average 65% relative abundance of temperate phage families at 3-4 months of age, interquartile range (IQR) [28.9%-95.9%]), shifting to a majority of virulent phage families at one year of age (average 50.7% [20.1%-85.2%] relative abundance of virulent phage families at 12 months of age, and 58.3% [36.3%-84.7%] at 24 months of age) (**Figure 3D**). The most abundant viral class was Caudoviricetes throughout all ages, which decreased in favor of Malgrandaviricetes after one year (**Figure 3E**). The relative abundance of several VFCs increased incrementally with age, in particular Gokushoviridae family number 84 and Almaviridae were of very low abundance in the earliest ages (3-6 months) before increasing after 6 months of age (**Suppl. Figure 5A**). In contrast, Flandersviridae and Alpaviridae family number 69 remained relatively stable in abundance across all ages. Viruses predicted to infect the Bacillota phylum were the most abundant throughout the sampling period. The viruses infecting the phyla Actinomycetota and Pseudomonadota showed a high abundance in the earliest samples (3-12 months) before decreasing, and Bacteroidota and Bacillota gained quickly in abundance before reaching a plateau around 12-18 months (**Figure 3F**). Viruses predicted to infect the genera *Bifidobacterium, Veillonella* and *Escherichia* were the most abundant in the 3-6 months samples, but steadily decreased in the following time periods (**Suppl. Figure**

5B). On the contrary, viruses infecting the genera *Alistipes*, *Faecalibacterium*, *Lactococcus* and *Blautia* were of low abundance in the 3-6 months samples but steadily increased in the following time periods.

Of the two human-infecting viruses *Anelloviridae* and *Adenoviridae* identified in the samples, *Anelloviridae* showed a high prevalence with varying abundance, even from the earliest samples, whereas *Adenoviridae* exhibited a modest prevalence with low abundance (**Suppl. Figure 5C,D**).

Figure 3. Virome development in early life.

A) Observed richness and B) Shannon diversity index of viral families (VFCs) from age 3-36 months. The points represent individual samples, and the lightly colored lines connect the samples of individual participants. The points and lines are colored by the age at sampling. The black line represents the LOESS locally smoothed regression fit with 95% confidence intervals (±1.96 standard errors). C) Principal coordinate analysis of Aitchison distances between samples quantified at the VFC level. The points represent individual samples, and the lightly colored lines connect the samples of individual participants. The points and lines are colored by the age at sampling. D) The abundance of phage families colored by their lifestyle, E) viral class, or F) predicted bacterial host phylum over age. The points represent the abundance of individual VFCs for each individual sample. The lines represent the LOESS locally smoothed regression fit with 95% confidence intervals (±1.96 standard errors).


A
Observed richness of VFCs
Observed richness
Age (months)
B
Shannon diversity index of VFCs
Shannon diversity index
Age (months)
C
Aitchison distance of VFCs
PCo2 (8.8%)
PCo1 (13.3%)
D
Lifestyle
VFCs abundance
Age (months)
E
Viral class
Viral class abundance
Age (months)
F
Viral host phylum
Viral host abundance
Age (months)
Age (months) (A,B,C)
Lifestyle (D)
Temperate
Virulent
Eukaryotic virus
Viral class (E)
Caudoviricetes
Malgrandaviricetes
Arfiviricetes
Viral host phylum (F)
Bacillota
Bacteroidota
Actinomycetota
Pseudomonadota

213

Having older siblings and birth mode associated with the early-life virome composition

With the rich longitudinal sampling and extensive information on participant characteristics we wanted to investigate the development of the observed VFC richness over time and its relationship with key environmental variables. Generalized additive mixed models (GAMM) were computed, enabling estimation of non-linear associations of age and other covariates with the observed richness, while controlling for the nested case-control study design and longitudinal sampling with random effects. This revealed that having older siblings was associated with a higher observed richness of the virulent VFCs in the earliest samples, with the effect tapering off with age (**Figure 4A** and **Suppl. Figure 6A**, effect estimate (exp(β)) = 10.1% higher richness, 95% confidence interval (CI) [3.6%;17%], p = 0.002, GAMM, **Suppl. Table 2**). Additionally, being born by cesarean section (C-section) was associated with a lower richness of the VFCs and virulent VFCs, with the strongest effect in the earliest samples that tapered off over time (**Suppl. Figure 6B,** VFCs -5.8% [-11.1%;-0.2%], p = 0.042, GAMM). There was also evidence of a higher richness after gluten had been introduced in the child's diet (VFCs 7% [3.2%;11%], p = 0.0003, GAMM). There was no evidence of a seasonal effect when using a cyclic smooth for sampling date (effective degrees of freedom (edf) = 1.03, $\chi^2 = 2.1$, p = 0.20). In a subset of the samples ($N_{individuals}$ = 96, $N_{samples}$ = 1,679) information about gastroenteritis events and antibiotics use during the sampling period was available, and a subanalysis revealed no association between the cumulative gastroenteritis rate or antibiotics course rate with observed richness (VFCs gastroenteritis rate p = 0.958, antibiotics course rate p = 0.331, GAMM).

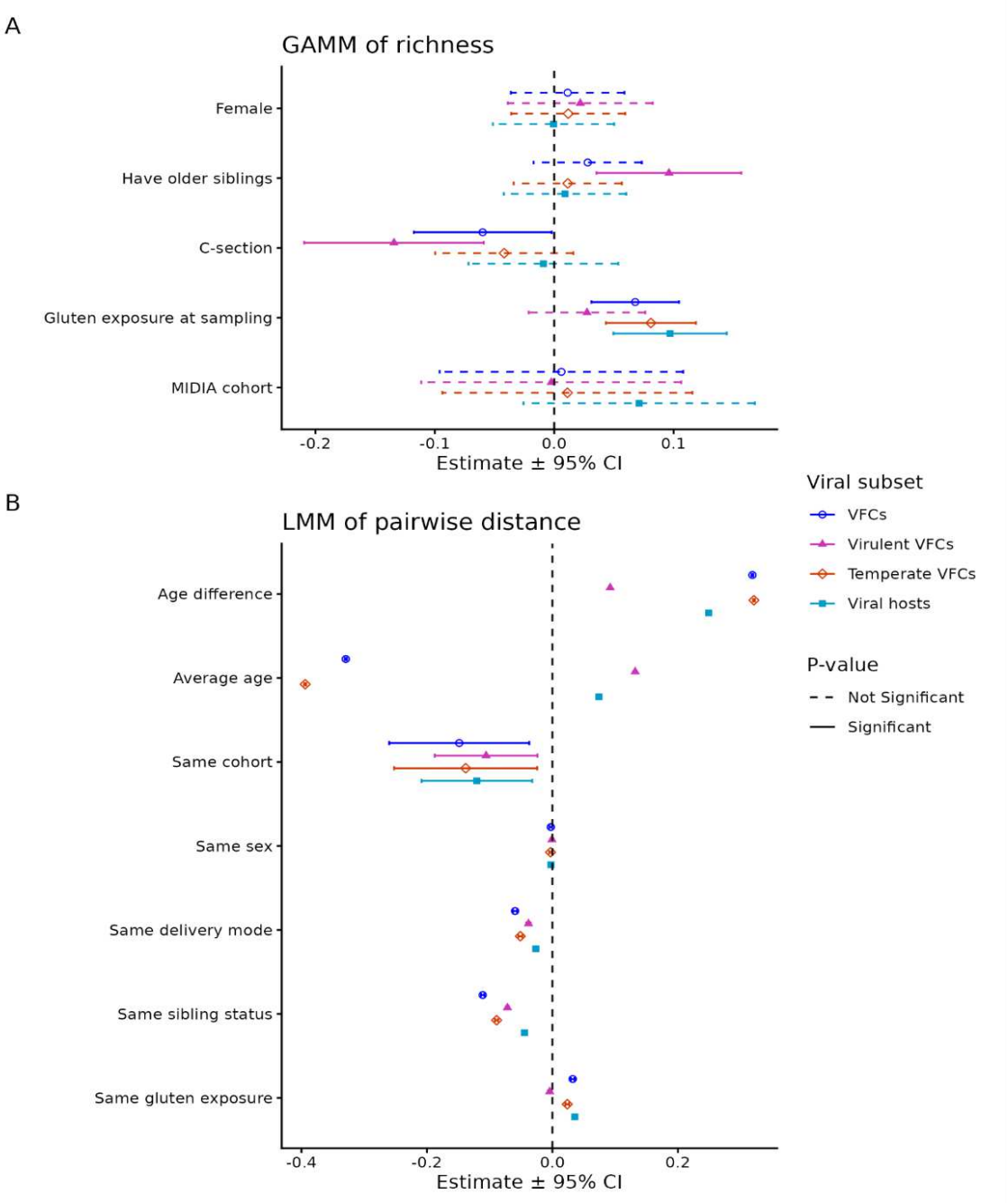
In addition to investigating the VFC richness, we also wanted to analyse the associations of environmental variables with the development of the overall community composition of the virome in early life (beta diversity). To this end, we conducted a permutational analysis of variance (PERMANOVA) of the Aitchison VFC distance (**Suppl. Figure 7A**, **Suppl. Table 5**). This revealed that the interindividual variation among children was the factor that explained the most variance in the beta diversity (24.7%, p-value cannot be accurately defined, **Suppl. Figure 7A** and **Suppl. Figure 8**), followed by age (2.8%, p = 0.001), and having older siblings (1.3%, p-value cannot be accurately defined).

When applying a PERMANOVA on longitudinal samples, permutations must be stratified by participant ID, and thus it cannot be used to test the significance of constant variables (e.g. sex or birth mode). As such, to assess the association of constant variables with the overall virome composition, we instead calculated the pairwise Aitchison distance between all pairs of children and applied a linear mixed model (LMM) to control for both individuals as random effects. This analysis allowed us to test the association between the pairwise distance between participants as the dependent variable and any constant variable as the

independent variable, while adjusting for covariates as well as the nested and longitudinal study design. This revealed that there was a larger pairwise Aitchison distance (indicating more different composition in a given pair) if the age difference between the samples in the pair were larger (**Figure 4B**, VFCs, $\beta = 0.32$ 95% CI = [0.32;0.32], $p<2x10^{-16}$, LMM, **Suppl. Table 3**). This was found for all viral subsets. On the contrary, if a given pair of children were on average older, their compositions were more similar in all viral subsets (VFCs, -0.33 [-0.33;-0.33], $p<2x10^{-16}$), except for the virulent VFCs (0.13 [0.13;0.13], $p<2x10^{-16}$). Samples belonging to children of the same cohort (VFCs, -0.15 [-0.26;-0.04], $p = 0.01$), children born by the same delivery mode (VFCs, -0.06 [-0.06;-0.06], $p<2x10^{-16}$), and children with the same sibling status (VFCs, -0.11 [-0.11;-0.11], $p<2x10^{-16}$) shared more similar compositions. In the subset of the samples with information about antibiotics use, we found that children who had received a cumulative larger number of antibiotic courses (adjusted for the age difference between their samples) had a more similar composition in the temperate VFCs (-0.10 [-0.18;-0.01], $p = 0.02$).

Figure 4. Associations between biological and technical factors and the alpha and beta diversity of the virome.

A) Fixed effects from the GAMM on observed richness. The random effects included in this model are: age, scaled log10 proportion of viral reads, sampling year, matching group, participant, and sequencing run. B) Fixed effects from the LMM on the average pairwise Aitchison distance. The random effects included in this model are: matching group, participant, and sequencing run. The color of the lines represent the viral subsets, and the dashed lines represent non-significant associations while the full lines represent significant associations. GAMM = Generalized additive mixed model. LMM = Linear mixed model.



The infant gut virome composition is modestly associated with the risk of developing CeD

Having described the development of the gut virome and identified its most influential early-life factors, we next assessed whether this development differed between children who later developed CeD and their controls. The observed richness development across early childhood was not significantly different among the children who developed CeD compared to controls (**Figure 5C**, VFCs p=0.699, GAMM adjusted for all covariates, **Suppl. Table 2**), similarly months to CeD seroconversion showed no independent linear association (p=0.518), nor an interaction effect with CeD status (interaction p=0.370).

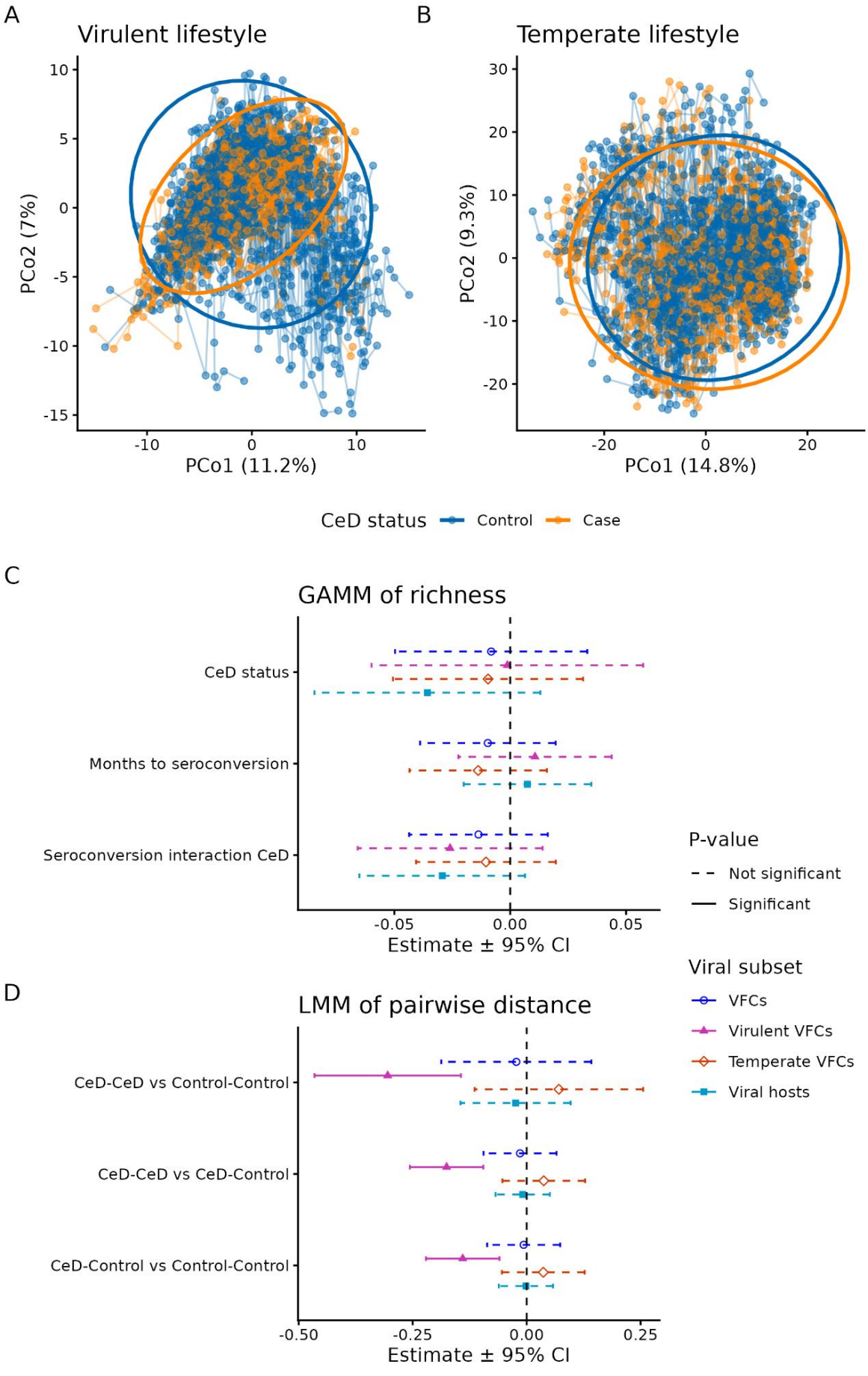
Moreover, the acquisition rates of novel VFCs over time were compared between children who developed CeD and controls (newly detected VFCs for each consecutive sample were quantified and divided by sampling interval, **Suppl. Figure 4E**). The acquisition dynamics of novel VFCs were not dependent on CeD status or months to seroconversion in any of the viral subsets (VFCs p=0.711, LMM adjusted for all covariates). Next, the rate of change in alpha diversity (step size) between consecutive samples within a participant over time was compared between children who developed CeD and controls (**Suppl. Figure 4F**). This analysis revealed no significant association in any of the viral subsets (VFCs p=0.142, LMM adjusted for all covariates), suggesting that the rate of change of alpha diversity did not differ based on future development of CeD or months to seroconversion.

Next, we wanted to investigate whether the overall virome composition differed between children who developed CeD and control children. First, we analysed the within-group pairwise Aitchison distances between CeD-CeD child pairs vs. the pairwise distance between control-control child pairs. This revealed a significant association only for the virulent VFCs (β = -0.31 [-0.47;-0.14], p = 2.5×10^{-4}), but no association for the other viral subsets (p>0.05), indicating that the children who later develop CeD have a more similar virulent virome composition among themselves, than the control children have among themselves. To further understand this relationship, we additionally analysed the between-group pairwise distance, by comparing the pairwise distances between CeD-CeD child pairs vs. the distances between CeD-control child pairs, and similarly by comparing the distance between control-control child pairs vs. the distance between CeD-control child pairs. This revealed a significantly more similar virulent virome composition among CeD-CeD children compared to the composition among CeD-control children (**Figure 5A, B, D**, β = -0.18 [-0.26;-0.1], p = 2.8×10^{-5}, LMM of the pairwise Aitchison distance comparing CeD-CeD vs. CeD-control pairs, adjusted for all covariates, **Suppl. Table 3**). There was no significant difference in any of the other viral subsets (p>0.05). Similarly, the pairwise distance of the virulent VFCs among CeD-control child pairs compared to control-control child pairs, was also significantly lower (β = -0.14 [-0.22;-0.06], p = 0.0008).

Next, we tested the stability of the developing virome between CeD cases and controls using two metrics: average distance to individual centroid, capturing stability across all samples, and the dissimilarity between consecutive samples within a child, capturing short-term stability over time. None of these metrics differed between CeD cases and controls (**Suppl. Figure 7C,D**, p=0.291 and p=0.371, respectively, **Suppl. Tables 6, 7**).

Figure 5. Virome development in CeD cases and controls.

A) Principal coordinate analysis of the Aitchison distance of the VFCs with a virulent or B) temperate lifestyle, at the viral family level. The points represent individual samples, and the colored lines connect the samples of individual participants. The color of the points and ellipses represent CeD case or control status. The ellipses represent the 95% data coverage for the CeD case and control groups. C) Fixed effects from the GAMM on observed richness. D) Fixed effects from the LMM on the average pairwise Aitchison distance. The color of the lines represent the viral subsets, and the dashed lines represent non-significant associations while the full lines represent significant associations. The random effects included in these models are: age, scaled log10 proportion of viral reads, sampling year, matching group, participant, and sequencing run. PCo = Principal coordinate. GAMM = Generalized additive mixed model. LMM = Linear mixed model. CeD = Celiac disease.


A
Virulent lifestyle
PCo2 (7%)
PCo1 (11.2%)
B
Temperate lifestyle
PCo2 (9.3%)
PCo1 (14.8%)
CeD status
Control
Case
C
GAMM of richness
CeD status
Months to seroconversion
Seroconversion interaction CeD
Estimate ± 95% CI
P-value
Not significant
Significant
Viral subset
VFCs
Virulent VFCs
Temperate VFCs
Viral hosts
D
LMM of pairwise distance
CeD-CeD vs Control-Control
CeD-CeD vs CeD-Control
CeD-Control vs Control-Control
Estimate ± 95% CI

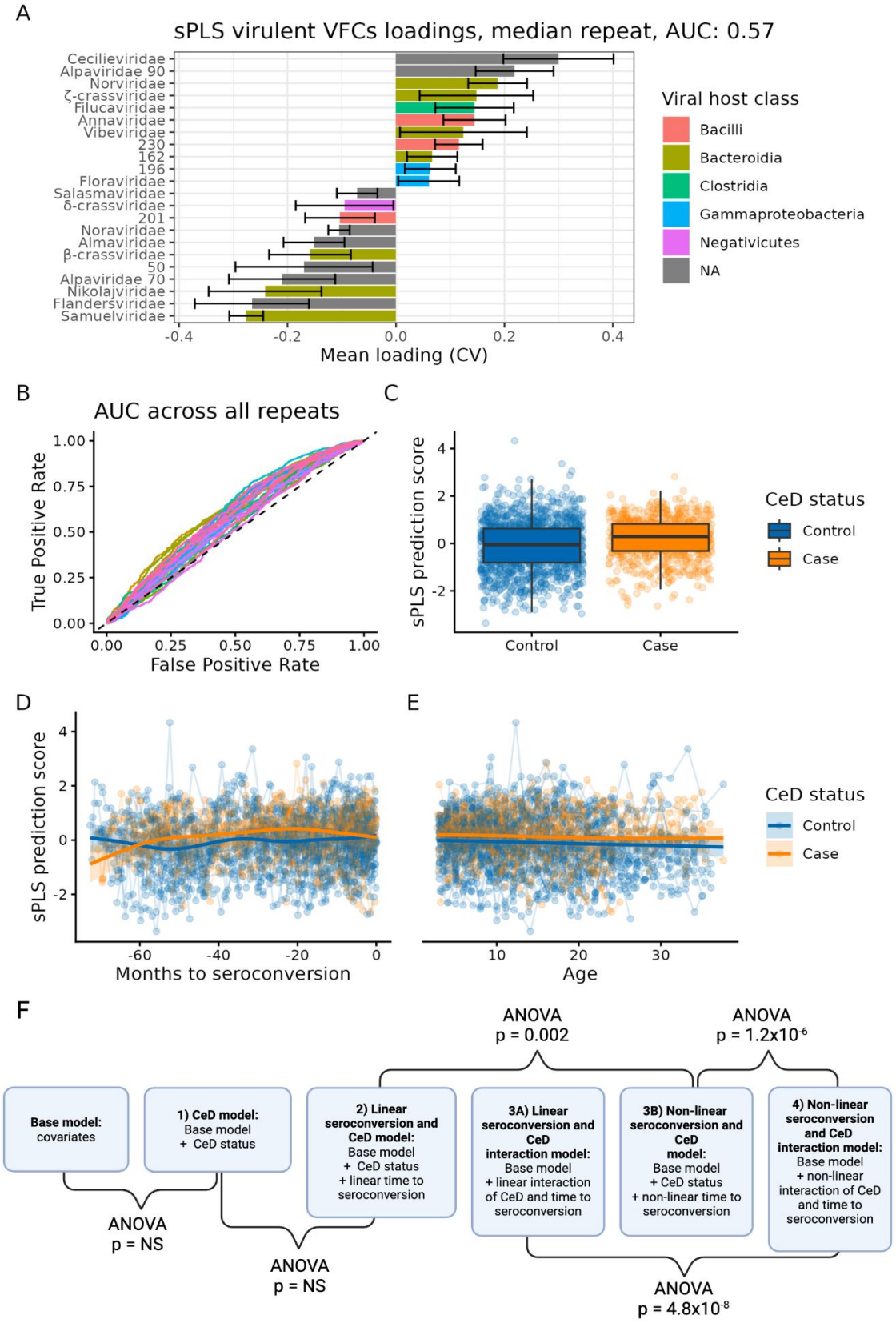
334

Virus families with a virulent lifestyle associated modestly with the risk of developing CeD

Next, we sought to express a maximally discriminative composite score based on viral family abundances, using a supervised learning model. We trained a sparse partial least squares (sPLS) model on virulent VFCs, which was able to moderately discern between children who developed CeD compared to controls (**Figure 6,** median test AUC across 31 repeats = 0.57, **Suppl. Table 4**). There was no association in alternative models based on all VFCs, temperate VFCs, nor the predicted bacterial host genus abundances (median test fold AUC = 0.53, 0.52 and 0.52, respectively). The sPLS model was naive to the longitudinal study design and covariates, so to account for the repeated sampling and nested case-control design of the study, the sPLS prediction scores were extracted (the probability that a sample belongs to a CeD case) and modelled as the dependent variable in a progression of GAMM. The sPLS prediction score can be interpreted as the predicted probability that a sample belongs to a CeD case. The GAMM was implemented to predict the sPLS score, while adjusting for covariates accounting for the study design. A series of GAMM were implemented in a progressive approach to analyse the effect of potential linear and non-linear associations between CeD status and months to seroconversion (**Figure 6F**). These analyses revealed no significant association between the virulent VFC sPLS score and CeD (ANOVA chi-squared test $p>0.05$), nor with a linear fit of months to seroconversion ($p>0.05$), or an interaction between the two (CeD * months to seroconversion, $p>0.05$). Nevertheless, when allowing for a non-linear fit of months to seroconversion to the virulent VFC sPLS score the association was significant (analysis of variance (ANOVA) model 2 vs model 3B, $p = 0.002$, **Figure 6F, Suppl. Figure 9**). Furthermore, allowing this non-linear fit of months to seroconversion and the sPLS score to differ between children who develop CeD and controls significantly improved the model fit, compared to a shared non-linear fit (ANOVA model 3B vs model 4, $p = 1.2 \times 10^{-6}$). This non-linear relationship is visualized per matching group in **Supplementary Figure 9**.

Figure 6. Machine learning prediction model (sPLS) of virulent VFCs predicting CeD status.

A) Plot of the mean loading for each of the virulent VFCs selected by the sPLS to predict CeD. The repeat with the median AUC was extracted and visualized. Only variables with a loading >0.05, and a standard deviation of the loading > mean loading are shown. The prediction score was extracted for this model and used as the dependent variable in a GAMM to test association with CeD, while accounting for the matched and longitudinal design by random effects for individuals nested within matching groups, and adjusting for covariates as fixed effects. B) AUROC curve of the sPLS model of virulent VFCs for all repeats. C) Boxplots demonstrating the sPLS predicted score from the median performing repeat, based on AUROC. The points represent the predicted score for each sample. The colors represent the CeD status. The box represents the 25th and 75th percentiles, the middle line represents the median, and the upper and lower whiskers represent the 5th and 95th percentiles. The black points represent outliers. D) The observed sPLS prediction score from panel A is plotted over months to seroconversion and E) age. Each point represents the observed prediction score for a sample, and the color refers to the CeD status. The thick lines represent the predicted sPLS scores from a GAMM with 95% confidence intervals (±1.96 standard errors). F) Overview of GAMM progressive testing for predicting virulent VFC sPLS score. The base model predicts the sPLS score as dependent variable with the following variables as independent variables (hereafter referred to as covariates): binary or linear: sex, having older siblings, cohort, sequencing depth, sampling month, delivery mode, gluten exposure at sampling, non-linear smooths: age, proportion of viral reads, and sampling year, random effects: matching group, participant, and sequencing run. All models that follow are based on the base model (all covariates), added with 1) CeD status 2) CeD status and linear time to seroconversion 3A) linear interaction of CeD status and time to seroconversion, 3B) CeD status and non-linear time to seroconversion, and 4) non-linear interaction of CeD status and time to seroconversion. The curly brackets indicate which models were compared with an ANOVA Chi-squared test. If p=NS the GAMM could not be formally tested by ANOVA because the resulting deviance difference was negative. sPLS = sparse partial least squares. VFC = viral family clade. AUC = area under the curve. GAMM = generalized additive mixed model. CeD = celiac disease. AUROC = area under the receiver operating characteristic curve. ANOVA = analysis of variance.


A
sPLS virulent VFCs loadings, median repeat, AUC: 0.57
Cecilieviridae
Alpaviridae 90
Norviridae
ζ-crassviridae
Filucaviridae
Annaviridae
Vibeviridae
230
162
196
Floraviridae
Salasmaviridae
δ-crassviridae
201
Noraviridae
Almaviridae
β-crassviridae
50
Alpaviridae 70
Nikolajviridae
Flandersviridae
Samuelviridae
-0.4
-0.2
0.0
0.2
0.4
Mean loading (CV)
Viral host class
Bacilli
Bacteroidia
Clostridia
Gammaproteobacteria
Negativicutes
NA
B
AUC across all repeats
True Positive Rate
False Positive Rate
C
sPLS prediction score
Control
Case
CeD status
D
sPLS prediction score
Months to seroconversion
E
Age
F
Base model: covariates
1) CeD model: Base model + CeD status
2) Linear seroconversion and CeD model: Base model + CeD status + linear time to seroconversion
3A) Linear seroconversion and CeD interaction model: Base model + linear interaction of CeD and time to seroconversion
3B) Non-linear seroconversion and CeD model: Base model + CeD status + non-linear time to seroconversion
4) Non-linear seroconversion and CeD interaction model: Base model + non-linear interaction of CeD and time to seroconversion
ANOVA p = NS
ANOVA p = NS
ANOVA p = 0.002
ANOVA p = 1.2x10-6
ANOVA p = 4.8x10-8

393

Discussion

Here, we describe the development of the gut virome from 3 months until 3 years of age in children who developed CeD and controls, based on more than 2000 fecal viromes collected from 117 children selected from two prospective birth cohorts. The virome was highly personalized, rapidly increasing in richness over the first years of life, while the composition underwent marked compositional shifts. The composition (beta diversity) was modified by environmental factors, including older siblings, birth mode and antibiotic treatments. The children follow similar development trajectories, nonetheless, we found differences in the viruses with a virulent lifestyle in children who develop CeD compared to controls.

The gut virome develops rapidly in the first three years of life, highlighted by the tremendous increase of unique virus entities from 3-18 months of age, and the extensive change in the overall virome composition throughout the entire age span studied. This is largely mirroring the bacteriome development at this age[35–37]. Virus families exhibiting a temperate lifestyle dominated the early samples, but decreased in relative abundance while virulent virus families started dominating around one year of age. This is in line with previous studies, suggesting that the initial seeded viruses originate from pioneer bacteria[6]. The bacteriome maturation in early life that has been described previously is also mirrored in the virome by their predicted viral hosts. The earliest samples contain mainly viruses infecting bacterial genera belonging to the Bacillota (formerly Firmicutes) and Actinomycetota (mainly represented by the family Bifidobacteriaceae) phyla. The Actinomycetota-infecting viruses decrease in relative abundance until approximately 18 months of age, along with a decrease in Pseudomonadota (Proteobacteria) -infecting viruses, while the viruses infecting Bacillota and Bacteroidota steadily increase in relative abundance over time. In the earliest samples, Caudoviricetes was the most abundant viral class, though around one year of age it decreased slightly in relation to an increase in relative abundance of Malgrandaviricetes. Our findings highlight the non-linear trajectory of the early-life gut virome developmental dynamics which largely validates what has previously been reported by others[22]. Yet, some early studies found that the observed richness of the gut virome was initially high and decreased with age[21,27,38], which is contradictory to what we demonstrate here and what several other studies have reported[27,38].

Firstly, we investigated the impact of several environmental factors on the within-individual gut virome composition (alpha diversity) in early-life. In line with previous studies, we did not find any significant associations of sex on viral diversity. Conversely, having older siblings was associated with a clear increased viral richness for the virulent phage families, which mirrors the general increase in bacteriome diversity observed elsewhere in children with

siblings[39,40]. Being born by C-section was associated with a decreased viral family richness, particularly for the virulent phage families. This is in line with what Zeng et al. previously reported on their bulk metagenome-derived viromes, that vaginally-born children had a higher richness[22], which was also driven by mainly virulent phage families, despite the difference in methods used. Further, an increase in the viral richness of the temperate VFCs and predicted viral hosts was observed after the introduction of gluten, though it is outside the scope of this study to discern whether this could be attributed to the introduction of any solid foods. Of note, we saw no particular associations of the cumulative number of gastroenteritis events or antibiotics courses, nor of the cohort, sequencing depth, or sampling month, on the viral richness in this age span.

The variance in the overall virome composition (beta diversity) was largely explained by the individual, followed by age at sampling. This demonstrates how unique the virome composition is, and highlights the tremendous maturation of the virome in early life, and the importance of considering age when investigating the virome in children. These results support previous findings of individual and age as strong drivers, such as the closer resemblance of the gut virome between unrelated infants as opposed to a child's own mother[22,23]. Interestingly, having older siblings was another strong predictor of the virome composition, but also the average use of antibiotics over the sampling period was associated. In contrast, delivery mode, cohort, and the number of gastroenteritis events only exhibited modest associations, which is in line with previous findings on the bacteriome in early life[41]. Similarly, Zeng et al. did not detect any associations of birth mode on their VLP virome composition, though they reported that it had a relatively large influence on the bulk metagenome-derived virome composition in the samples collected at birth (3.8% variance explained), which decreased abruptly already at 1 month of age[22]. Among the technical variables, sequencing depth, and sequencing run were of great importance, though sampling date had a negligible effect.

The observed bacteriophage richness increased over time along the same trajectory for both cases and controls, which echoes previous literature on the bacteriome in CeD[19,42–44]. They also had a similar overall virome composition in terms of beta diversity, when comparing the average distance to an individual's centroid between cases and controls. Furthermore, we tested whether short-term changes in the virome composition were associated with the development of CeD by an analysis of the stepwise sample-to-sample change in the virome development for each child. This analysis did not suggest any differences between cases and controls, which is in contrast to what we have previously found for the human-infecting viruses in these children, where infections with parechovirus and enterovirus were more frequent before CDA seroconversion[7,17,18].

We investigated whether the pairwise dissimilarity between samples differed among children who developed CeD. Here, we found a significantly more similar composition in the virulent virus families among CeD cases when compared to both control-control compositions or CeD-control compositions. Moreover, we employed an sPLS model based on virulent virus families which could modestly discriminate CeD cases from controls. This was expressed as a compositional score of higher abundance of Cecilieviridae and Alpaviridae (family number 90), and lower abundance of Samuelviridae and Flandersviridae in cases compared to controls. When accounting for the study design and covariates in a GAMM, a significant non-linear association between the sPLS model prediction score and CeD status following the seroconversion time-scale became apparent, suggesting that the abundance of the selected viruses are indicative of the risk of CeD before onset of disease. The virus families selected by the model were not infecting a particular bacterial host, but rather seemed to describe an overall virulent virome signature that differed between cases and controls.

Published data on the role of the gut virome in the development of CeD are limited. In 2019, the TEDDY international consortium analysed shotgun metagenomic sequencing data from children's stools prior to the development of CDA[45]. The viromes of 83 case-control pairs were analysed with a specific focus on enterovirus, adenovirus, astrovirus, norovirus, reovirus and rotavirus. An inverse association with subsequent CDA was observed with the cumulative exposure to the above viruses under one year of age, whereas in the time window between 1 and 2 years of age, cumulative exposure to any of the above viruses, and to enterovirus in particular, was associated with an increased risk of CDA. This finding is consistent with our previous study of the Norwegian MIDIA cohort[18], including the interaction between enterovirus and gluten exposure. Of note, Lindfors et al. used a tool primarily aimed at detecting human viruses[46], leaving phages outside the scope of their investigation. Another study using metagenomic sequencing in CeD has been published by a group from Saudi Arabia: it had a cross-sectional design and examined the faecal and gut mucosal viromes of 40 children with newly diagnosed CeD and 39 controls[47]. They found numerous taxonomic entities associated with the condition, however this study has considerable limitations, such as the cross-sectional design that is unable to test any temporal associations and is prone to reverse causation, as well as the use of unsuitable statistical methods and potentially biased selection of controls.

A major strength of this study is the nested case-control study that was established within the two international prospective birth cohorts DIPP and MIDIA. This supports the findings reported here as being of universal nature, thus not specific to a single cohort. The participants were closely followed throughout the first years of life or until seroconversion to positivity for CeD autoimmunity, with frequent sample collection enabling discovery of

differences manifesting before disease onset. Several environmental variables that can be highly relevant for the incidence of CeD as well as the microbiome composition and maturation, such as birth mode, gastroenteritis infections, antibiotics use, and siblings, were collected and analysed in this study. There is a lack of prospective cohorts in this field: previous studies either lacked information on phages, or compared controls to already CeD diagnosed individuals, which bears the risk of reverse causation. As the results presented here are based on samples from a longitudinal nested case-control study design, the analyses are ideally applicable to describe the maturation of the virome, and identify CeD-associated virome changes in the early window of opportunity for risk-reduction and prevention, before onset of disease. Another crucial strength of the study is the virome sample processing in concordance with the database-independent computational approach cataloging the so-called viral dark matter, allowing us to investigate more than 5,000 novel vOTUs and 263 novel VFCs, of both virulent and temperate lifestyles.

This study has several limitations. Firstly, the participants were chosen based on HLA-associated risk of CeD, which limits the findings to a relatively large subset of the population; approximately 40-50% of the North European population carry these haplotypes. Previous studies suggest that the human genome, including HLA-DQ2, is modestly associated with the microbiome composition[48–50]. Secondly, this study only analysed the fecal DNA-virome, lacking the RNA-virome, though this was thoroughly investigated in our previous studies with specific reverse transcriptase polymerase chain reaction detection or specific antibodies[7,17,18]. In contrast, this study is also missing data on the bacteriome which, considering that the majority of the sequenced viruses infect bacteria, represent an important part of the ecology of this system.

This study investigates the development of the infant gut virome in two independent Nordic birth cohorts throughout the first three years of life, revealing a vast expansion in viral richness and describing the dynamics of the different viral classes and lifestyles over time. Several early-life factors were associated with observed richness and the virome composition in early life. We did not identify any association between future development of celiac disease (CeD) and viral richness in this age range. However, children who later developed CeD shared a more similar virome composition when compared to controls. Additionally, a machine learning model identified differences between cases and controls in the viral families with a virulent lifestyle, with further analyses suggesting a non-linear relationship between this composition and time to CeD seroconversion. Ultimately, this suggests a critical window of opportunity for targeted interventions of virulent viral families to reduce risk or prevent the onset of clinical CeD.

Online Methods

Study population

We established a case-control study nested in two birth cohorts, Norwegian MIDIA[32] and Finnish DIPP[31], recruiting infants with HLA-DQ and -DR variants associated with susceptibility for type 1 diabetes (T1D), which also confer susceptibility to CeD. Cases developed tTGA (autoantibodies were detected in the blood samples) during their childhood and were subsequently diagnosed with CeD (clinically determined based on the established 1990 or 2012 ESPGHAN criteria). Two (or in some cases one) tTGA- and CeD-free controls from the same cohort were matched per case by date of birth, place of birth, and duration of follow-up at least to the age when the case was diagnosed.

The biobanked faecal samples, collected from 3 until 36 months of age, were used for virome profiling. In total, 2183 faecal samples were processed coming from 41 cases and their 76 matched controls. The MIDIA cohort provided 1330 samples from 74 individuals (25 cases with 49 matched controls). The DIPP study provided 853 samples coming from 43 individuals (16 cases with 27 matched controls). In order to avoid reverse causality, only stool samples before the tTGA development in cases and the respective control samples were included in the analysis. Both the MIDIA and the DIPP cohort were originally designed to study the etiology of T1D, commencing shortly after birth, all children provided regularly collected questionnaires, blood and stool samples, and their tTGA and CeD status were carefully evaluated later, as described below. Participant characteristics are depicted in **Table 1**.

MIDIA cohort

MIDIA participants (children born throughout Norway 2001-2007) were followed if they were identified at birth to carry the genotype associated with the highest HLA-associated risk of T1D (HLA-DRB1*04:01-DQA1*03-DQB1*03:02/DQA1*05-DQB1*02). This genotype consists of two haplotypes conferring also the highest risk for CeD, termed HLA-DQ2.5 and HLA-DQ8. Monthly stool samples (during 3-36 months of age), plasma samples (at 3, 6, 9 and 12 months of age and then annually) and regular questionnaires were collected. These questionnaires recorded for example the intake frequency of selected food items, infections, symptoms, doctor contacts, as well as use of antibiotics (duration and timing)[18,51]. During 2014-2016, the active MIDIA participants were approached with an offer to participate in CeD disease testing, of which 220 agreed and underwent CeD screening and clinical evaluation. Their most recent blood samples were tested for tTGA (EliA Celikey IgA/EliA, Thermo Fischer Scientific, Phadia AB; Uppsala, Sweden), furthermore, 13 children had already been diagnosed with CeD. In total, 25 CeD cases were enrolled in the present substudy, based on the ESPGHAN 2012 criteria[17]. Each case was matched for date of birth, county of residence and follow-up duration with one to two tTGA and CeD-free controls. By retrospective analysis of biobanked plasma samples, the last tTGA negative plasma sample of each case was determined; six cases of 25 seroconverted after the end of stool sampling. All stool samples taken before the tTGA seroconversion interval from 25 cases and their 49 matched controls were subjected to DNA virome analysis. The last date for determining T1D status (via linkage to the Norwegian childhood clinical diabetes complete population coverage registry[52]) was in December of 2021. MIDIA fecal samples were processed as described by Cinek et al[53]. In brief, samples were suspended in phosphate buffered saline

with the addition of antibiotics and fungicides, and samples were vortexed and centrifuged at 4000 g for 30 min. Supernatants were frozen at -80°C until preprocessing and nucleic acid extraction (described below). Bacterial pellets were discarded and consequently, bacteriome profiling from MIDIA samples was not possible.

DIPP cohort

The DIPP (Type 1 Diabetes Prediction and Prevention study) birth cohort[31] that began screening newborn infants from the general population in 1994 (born between 1997-2004), was also originally designed to study T1D. Within the DIPP cohort, a CeD sub-study was created, consisting of 16 cases with 27 controls matched for birth date and county of residence. DIPP children included in the present study had the following genotypes: HLA-DQB1*02/03:02, the *03:02/x genotype [x ≠ *02, *03:01 or *06:02] and boys with the genotype DQB1*02/y-DQA1*05/z [y ≠ *03:01, *03:02, *06:02, *06:03; z ≠ *02:01]. Serum samples were collected every 3–6 months between 3-24 months of age, and then 1-2 times annually. Stool samples were collected monthly from 3-24 months of age, and then annually. Their most recent blood samples were tested for tTGA (EliA Celikey IgA/EliA, Thermo Fischer Scientific, Phadia AB; Uppsala, Sweden). Additionally, 3-day food records at 3, 6, and 12 months and 2-, 3-, 4-, and 6-year visits were collected, alongside questionnaires regarding gastrointestinal infections and antibiotics intake. The last date for determining T1D status was in March of 2024 (diagnosis determined at representative pediatric departments of DIPP university hospitals based on World Health Organization and American Diabetes Association criteria). The biobanked DIPP stool samples were processed similarly to those from the MIDIA study, except they were stored at -80°C without the addition of antibiotics or fungicides, and centrifugation for supernatant separation was performed at 600 g for 30 min. Detailed protocol of DIPP stool sample processing is published by Salminen et al.[54].

Inclusion and ethics

MIDIA ethics committee approval: the Norwegian Regional Committees for Medical and Health Research Ethics Sør-Øst A, S-08164A MIDIA. DIPP study was approved by the Ethics Committee of the Wellbeing Services County of North Ostrobothnia, Oulu, Finland.

DNA extraction, library construction and massively parallel sequencing

We performed shotgun metagenomic sequencing of the DNA virome. The biobanked supernatant after low-speed centrifugation of stool samples was thawed and homogenised by vortexing, then re-centrifuged at 17,000 g for 3 minutes to collect cell debris, and the liquid phase filtered through centrifugation 0.8 micrometer filter. The QiaAmp Viral RNA mini kit (Qiagen, Hilden, Germany) was used for co-purification of DNA and RNA from the filtrate, with the addition of only a limited amount of carrier RNA. Libraries were prepared by a multi-step process starting by sequence-independent single-primer amplification (SISPA) in three independent reactions differing by the primer tag: the first and second strands of the random fragments were synthesized using Klenow fragment. Then the unused reaction components were digested by shrimp alkaline phosphatase and exonuclease I (both Thermo, US). Amplification using the tag primer followed. The products were then purified and size-limited by Ampure beads (Beckman Coulter). Each SISPA product was checked for the expected appearance (a continuous smear of DNA without noticeable bands).

The SISPA products were pooled by sample, and fragmenting and tagging was performed using the Nextera XT library construction kit (Illumina). Indexing with unique dual indices was performed by a limited count of PCR cycles, and purification followed. The ensuing libraries were equalised and sequenced on the NextSeq 500 instrument using NextSeq 500/550 Mid Output Kit v2, with the addition of 5 % of the *PhiX* control library to balance signal. The libraries were sequenced bidirectionally, at *2x*150 cycles. The detailed wet-lab protocol is publicly available[55] and performed with the highest specificity and sensitivity in a recent quality control round of metagenomic protocols[56].

Bioinformatics processing

An overview of the virome processing pipeline is described in **Suppl. Figure 2**. Virome libraries were sequenced on the NextSeq 500 platform to a mean depth of 2 Gb per sample with paired-end 2 x 150 bp reads for a total of 2,183 samples. Virome reads were quality filtered and trimmed using Fastq Quality Trimmer/Filter v. 0.0.14 (options -Q 33 -t 13 -l 32 -p 90 -q 13), and residual Illumina adaptors were removed using cutadapt (v2.0). Trimmed reads were deduplicated using the VSEARCH[57] v. 2.4.3 derep_prefix and then assembled with Megahit v. 1.2.9 separately for each sample. Contigs larger than 1000 bp from either cohort's samples were pooled and subject to pairwise alignments between each other using BLAT v. 35. Clusters of contigs that showed more than 95 % ANI were collapsed to the longest non-repetitive representative, at the cohort-level. Next, Genomad v. 1.5.1 was run on the contig representatives from each cohort, and contigs tagged as viral pooled across cohorts and subject to one more round of 95% ANI clustering to merge redundant contig representatives into vOTUs common to all three cohorts. Finally, the vOTUs were subject to a final stage of QC: Minimum length 2500 bp for inoviruses, microviruses and anelloviruses. Minimum length 10000 bp for caudoviruses and adenoviruses. In addition, caudoviruses needed to encode at least two proteins belonging to either the head or tail functional category. Encoded functional categories were obtained by first pooling all proteins detected by Genomad (Prodigal GV) from all vOTUs and subjecting them to an all-against-all protein sequence alignment using Fasta v. 36. Clusters of orthologous viral proteins were then made after filtering the all-against-all alignment with minimum coverage thresholds. Protein pairs that passed the thresholds were passed to MCL v. 14-137 which defined the viral ortholog (VOG) clusters. All members of each VOG were aligned in a multiple sequence alignment (MSA) for the entire VOG using MUSCLE v. 3.8.425. Each MSA was then searched against an HHblits formatted version of the Prokaryotic Virus Remote Homologous Groups (PHROG) database using HHsearch v. 3.8.425. The PHROG functional category was then transferred to the VOG if the match to the corresponding PHROG had a probability above 90%. The VOG and PHROG category information was further used to make visualisations of each viral family for visual confirmation. Viral protein sequences were embedded using the ProtT5 protein language model to obtain weighted mean protein embeddings for each vOTU. The viral lifestyle for each vOTU was determined by passing the vOTU embeddings through a linear classifier. Lifestyles were also determined at the VFC-level by mean pooling vOTU embeddings by VFC affiliation and running the same linear classifier. VFCs present in fewer than 30% of the samples were omitted from downstream analyses to filter possibly spurious viral taxa. This was done because viral families lack conserved marker genes and the correct approach for delineating them is still debated, hence the 30% prevalence cutoff ensured removal of spurious VFCs that may not represent real viral families (**Suppl. Figure 3**). Similarly, only the predicted viral host genera

that were detected in >10% of samples, were kept for downstream analyses (**Suppl. Figure 3**).

Statistical analysis

All analyses were carried out in R v. 4.2.0. Out of the 2,183 samples processed, 2,094 had matching meta-data and were collected from children under 40 months of age, and were used downstream for the analysis. Analyses were adjusted for technical and biological covariates: sequencing run ID, proportion of viral reads (log_{10} scaled), and sequencing depth (log_{10} scaled), sampling month, sampling year, age in months (scaled), study cohort, gender, having older siblings (yes/no), delivery mode by C-section (yes/no), gluten exposure at sampling (yes/no), antibiotics course rate (cumulative antibiotics courses until last fecal sample, divided by the total difference in time between first and last fecal sample), gastroenteritis event rate (cumulative events until last fecal sample, divided by the total difference in time between first and last fecal sample), participant ID, and matching group. In total, 14 participants had missing information about age of gluten introduction, which was imputed by the mean from their cohort. In total three samples had missing information about fecal sampling date and were excluded from permutational analysis of variance (PERMANOVA), generalized additive mixed models (GAMM) and linear mixed model (LMM) analyses, but kept for all other analyses. In a subset of the samples ($N_{individuals}$ = 96, $N_{samples}$ = 1,679) information about gastroenteritis events and antibiotics use was available, and subanalysis focusing on these two variables were included with this smaller sample size.

GAMM were fitted for analysing alpha diversity (described below the header *Alpha and beta diversity*) and sPLS scores (described below the header *Sparse Partial Least Squares (sPLS)*). For these models the following covariates were modeled as fixed linear or binary effects; study cohort, gender, having older siblings, sequencing depth, delivery mode by C-section, and gluten exposure at sampling. The following covariates were modeled with penalized regression splines (smooths); age, proportion of viral reads (with maximum of 15 basis functions k=15), sampling month (cyclic cubic basis function bs="cc", with maximum of 12 basis functions k=12) and sampling year. The following variables were modeled as random-effect smooths (bs='re') to account for the study design (nested matching groups and longitudinal sampling); participant, matching group, and sequencing run.

Alpha and beta diversity

Alpha diversity was calculated using the function estimate_richness from the *phyloseq* R package v. 1.54.0 on the raw counts of the VFCs and predicted viral host genus levels. Beta diversity was calculated using the *vegdist* function from the *vegan* package v. 2.7-2, with Aitchison distance on counts with a pseudocount of 1. PERMANOVA was performed with the function *adonis2* from the package *vegan* v. 2.7-2, with the settings by=margin and permutations=999, and the strata was defined to respect participant ID. For the variables that were variable within a participant (e.g. sampling date, age, months to seroconversion, sequencing run ID, sequencing depth, proportion of viral reads), the R^2 and p-values were reported, but for the variables that were constant within a participant (e.g. participant ID, CeD status, gender, number of older siblings, delivery mode, number of samples, and cohort), only the R^2 was reported, as the p-values cannot be accurately defined within the participant strata.

GAMM were fitted with the *gam* function from the *mgcv* package v. 1.8-40, with method=REML and select=TRUE. For observed richness, the negative binomial family was used. GAMM for CeD were tested by comparing a 0) "base model" with only the covariates (described in detail under the *Statistical analysis* header above), to several different GAMM 1) base model + CeD status (as a linear effect), 2) base model + CeD status + months to seroconversion (linear effects), 3) base model + CeD status linear interaction with months to seroconversion, and 4) base model + CeD status non-linear interaction with months to seroconversion. The following tests were done using an ANOVA with Chisquared test: model 0 vs model 1, model 1 vs model 2, model 2 vs model 3A, model 3A vs model 3B, model 3A vs model 4, and model 3B vs model 4.

Changes in alpha diversity across subsequent samples within individuals (alpha diversity step-size) were calculated by computing the difference in Shannon richness and observed richness of VFCs, discarding the first sample per individual to establish the initial time step. These step-wise changes in alpha diversity were used as the dependent variable in a LMM to estimate differences between cases and controls, using the *lmer* function from the package *lmerTest* v. 3.1-3. Months to seroconversion, age, sex, and cohort were included as covariates, alongside the time interval between samples and the total number of samples per individual. Random effects for participant ID nested within matching group were included to accommodate the longitudinal and matched nested study design.

The rate of new viral acquisition was calculated by tracking the presence/absence of vOTUs and VFCs along consecutive samples for each individual. For each sample, the number of viral taxa, novel viral taxa not previously observed in that individual, and the cumulative unique viral taxa count were calculated, along with the acquisition rate (novel viral taxa over time). Acquisition rate and cumulative novel taxa were used as the dependent variable in a LMM to estimate differences between cases and controls, using the *lmer* function from the package *lmerTest* v. 3.1-3. Months to seroconversion and age were included as covariates, alongside sex and cohort. Random effects for participant ID nested within matching group were included to accommodate the longitudinal and matched nested study design.

Pairwise beta diversity distance

Pairwise Aitchison distances were calculated between all samples. Sample pairs originating from the same participant were excluded. The pairwise distances were used as the dependent variable in a LMM using the *lmer* function from the package *lmerTest* v. 3.1-3. The effect of several technical and biological factors were tested as independent variables.

Continuous variables were encoded as the difference and the average between each pair of samples (age and sequencing depth). Sampling date was encoded as difference in sampling month and total difference in sampling date (including sampling year). Binary variables were encoded as being the same or different between each pair of samples (cohort, gender, birth mode, older siblings yes/no, gluten exposure at the time of sampling, sequencing run). Random effects for participant ID nested within matching group were included to accommodate the longitudinal and matched nested study design. Furthermore, we tested whether children who later developed CeD had a difference in their pairwise distances (CeD-CeD distance) compared to controls (control-control distance) *or* compared to CeD-control distance, while controlling for the mentioned covariates. Positive coefficients indicate greater between-participant dissimilarity, whereas negative coefficients indicate greater similarity.

Distance to centroid

For each participant, the distance of each sample to the centroid of that participant's samples was calculated based on the pairwise Aitchison distance using the *betadisper* function from the package *vegan* v. 2.7-2. This distance to centroid was used as the dependent variable in a LMM using the *lmer* function from the package *lmerTest* v. 3.1-3. The effect of several technical and biological factors were tested as independent variables (age and sequencing depth, cohort, gender, birth mode, older siblings yes/no, gluten exposure at the time of sampling, sampling year and sampling month). Random effects for sequencing run, and participant ID nested within matching group were included to accommodate the longitudinal and matched nested study design. Furthermore, we tested whether children who later developed CeD had a difference in their distance to centroid compared to controls. Positive coefficients indicate greater within-participant dispersion (i.e. larger distances to centroid for a participant), whereas negative coefficients indicate lower dispersion.

Beta diversity step size

Successive Aitchison distances between consecutive longitudinal samples from each participant were calculated to quantify the magnitude of temporal compositional change over time (step size). These successive distances were used as the dependent variable in a LMM using the *lmer* function from the package *lmerTest* v. 3.1-3. The effect of several technical and biological factors were tested as independent variables (age interval between samples, average age between the two consecutive samples, average sequencing depth, difference in sampling month and elapsed calendar time (including sampling year), cohort, gender, birth mode, older siblings, and discordant gluten exposure between consecutive samples. Random effects for participant ID nested within matching group were included to accommodate the longitudinal and matched nested study design. Positive coefficients indicate greater compositional change between consecutive samples, whereas negative coefficients indicate greater compositional stability over time.

Sparse Partial Least Squares (sPLS)

For the Sparse Partial Least Squares (sPLS), we used *mixOmics* v. 6.22.0 with a custom cross-validation design (https://github.com/jonathanth/caspoc). We randomized each matching group across 5-fold cross validation splits with 31 repeats for stability, choosing the best model based on median performance (AUC) across repeats in up to 3 components. For each iteration, 3 folds were used for training, 1 for tuning and 1 for testing. Model hyperparameters were chosen based on tuning folds. From this, only the viruses that fulfilled both of the following criteria were plotted: 1) the virus was selected in a majority of the folds (>3 folds), and 2) the virus mean loading was larger than the standard deviation. The median prediction across repeats for each sample was extracted and used downstream as the sPLS-score. Only the sPLS-score from the sPLS models with a median AUC>0.55 were used for downstream analysis. The sPLS-score was scaled before being used in the GAMM as the dependent variable, and the default family Gaussian with identity link function was used, with the parameter select=TRUE. In this GAMM we again adjusted for the previously mentioned covariates, and tested the effect of modeling CeD status and seroconversion with linear or non-linear interactions. GAMM for the sPLS-score were tested by comparing a 0) "base model" with only the covariates (described under the methods section *Statistical analysis* above), to several different GAMM 1) CeD status 2) CeD status and linear time to

seroconversion 3A) linear interaction of CeD status and time to seroconversion, 3B) non-linear CeD status and time to seroconversion, and 4) non-linear interaction of CeD status and time to seroconversion. The following tests were done using an anova with Chisquare test: model 0 vs 1, model 1 vs 2, model 2 vs 3A, model 3A vs 4, and model 3B vs 4.

Acknowledgements

We thank the children and families of the MIDIA and DIPP cohort studies for their support and commitment. We acknowledge and appreciate the unique efforts of the clinicians involved in the studies who followed up on the children, collecting samples and establishing diagnoses. We acknowledge and thank Kjersti S. Rønningen for planning and running MIDIA from its inception. Kjersti sadly passed away in 2025. We acknowledge Trond Rasmussen at NIPH for data management in MIDIA. MIDIA is supported by the Norwegian Institute of Public Health, and has received funding from the Research Council of Norway (Storforsk). This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 874864 (HEDIMED project). O.C. was financed by the project *National Institute of Virology and Bacteriology* (Programme EXCELES, ID Project No. LX22NPO5103) - Funded by the European Union - Next Generation EU. U.B. was supported by the Novo Nordisk Foundation (grant NNF20OC0061029), Beckett-Fonden, and is supported by the BRIDGE – Translational Excellence Programme (bridge.ku.dk) at the Faculty of Health and Medical Sciences, University of Copenhagen, funded by the Novo Nordisk Foundation under grant agreement no. NNF23SA0087869.All funding received by COPSAC is listed at https://copsac.com. The Lundbeck Foundation (R16-A1694), the Ministry of Health (903516), the Danish Council for Strategic Research (0603-00280B) and the Capital Region Research Foundation have provided core support to the COPSAC research centre.

Author contributions

UB, KC, SAS, HH, LCS, OC, and JTh conceptualized and designed the study. KC, GT, KH, MC, JLa, JLe, KK, MK, JTo, RV, KS, HH, LCS, and OC were responsible for acquisition of data. SAS performed all bioinformatics analyses and generated Figure 2. UB performed all statistical analyses and generated tables and figures with contributions from KC, SAS, OC, and JTh. UB, SAS, GT, KB, JS, LCS, OC, and JTh interpreted the data. UB drafted the manuscript with contributions from KC, SAS, LCS, OC, and JTh. All authors reviewed and agreed to the published version of the manuscript.

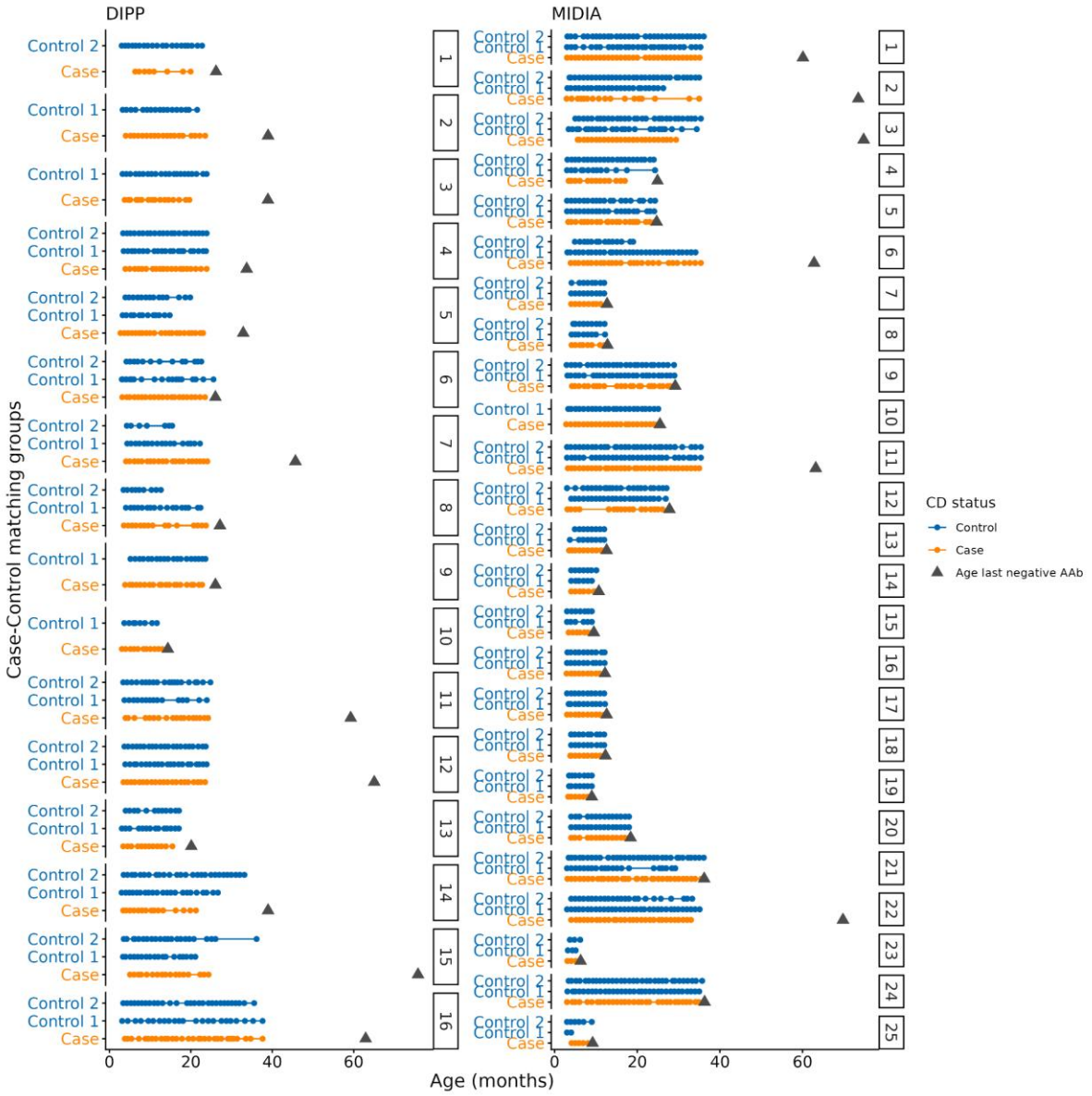
Competing interests

All authors declare no competing interests.

Supplementary Figures

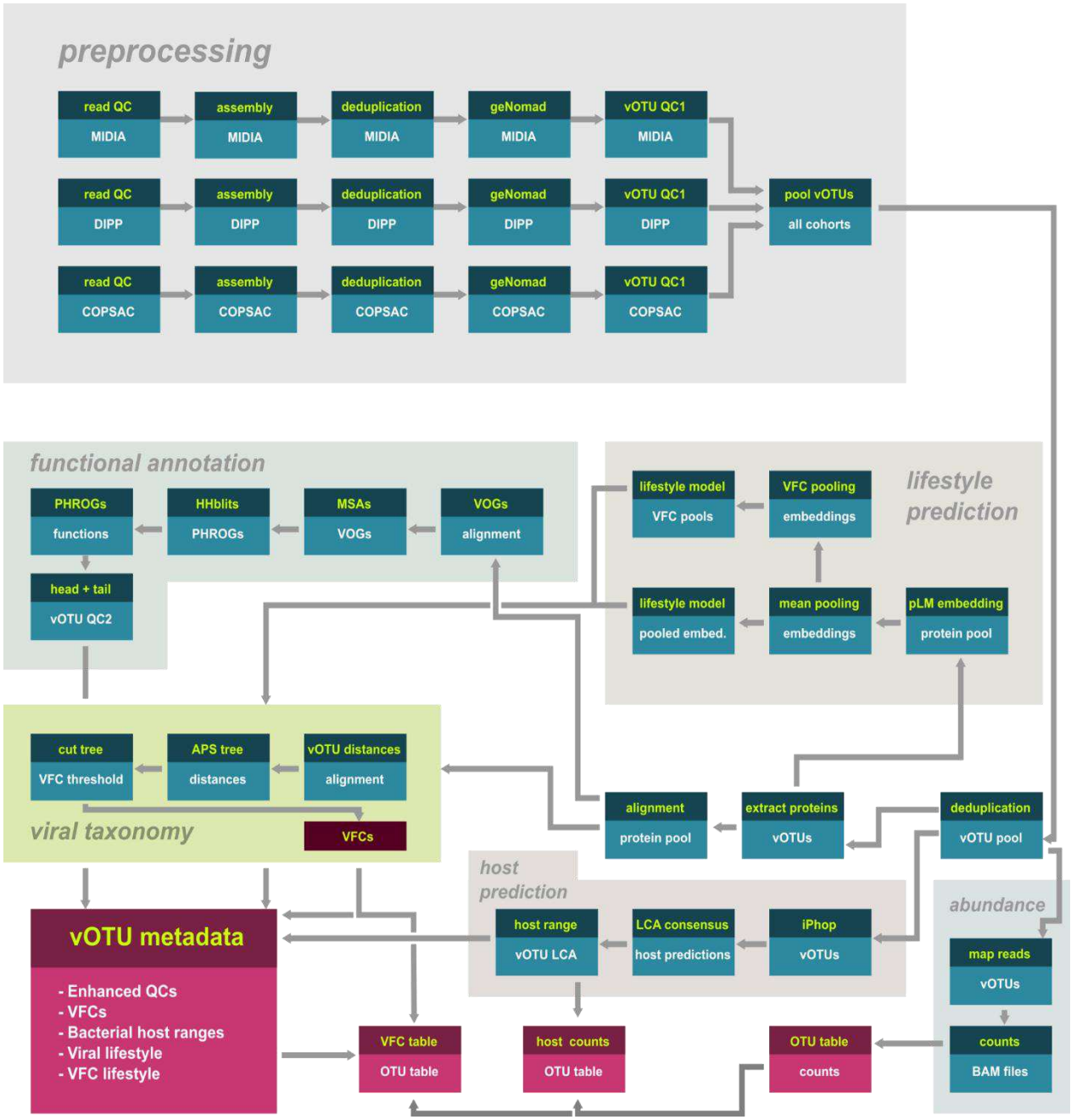
Supplementary Figure 1. Sample overview.

The two columns represent the DIPP and MIDIA cohorts, respectively. Each row represents a participant, with points representing their samples over age. The color indicates CeD status. The dark grey triangles indicate the age when the last blood sample was collected before autoantibodies were detected in the cases' blood samples. The grouping of the rows, numbered on the right, represents matching groups, with a single case and either one or two control participants per group.



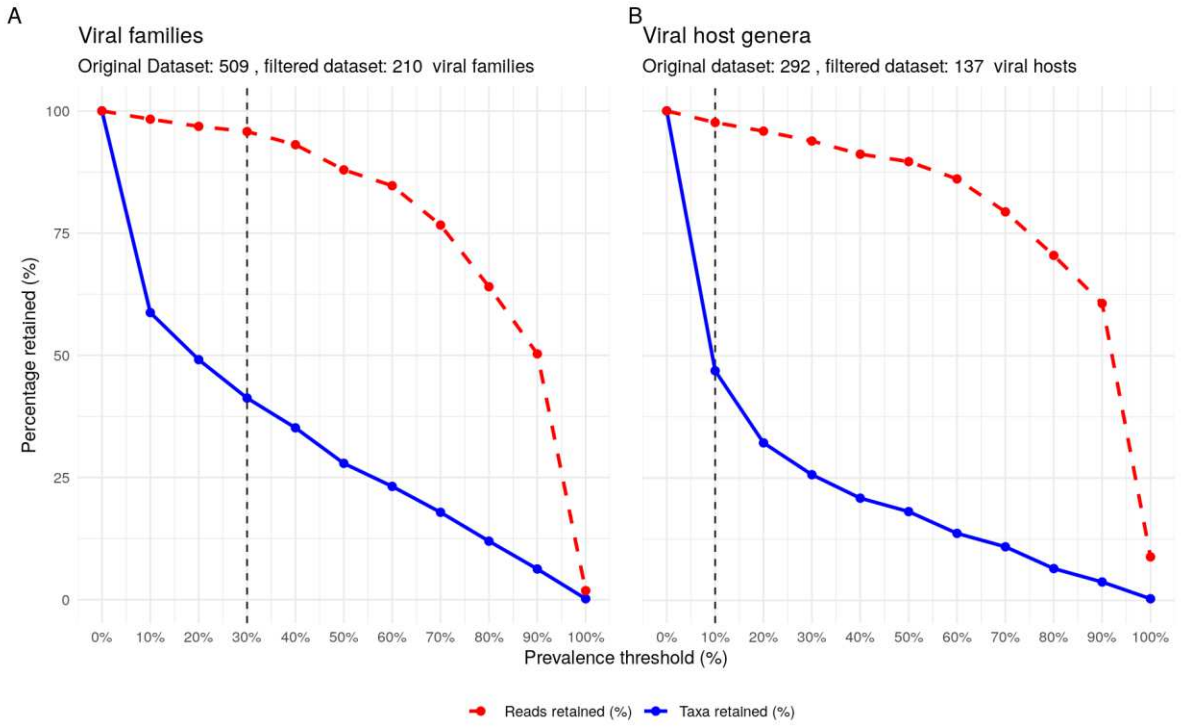
Supplementary Figure 2. Overview of virome processing pipeline.

Sequenced reads underwent quality control and were assembled before deduplication across samples at the species-level, separately per cohort. geNomad was used to flag the viral species, which were then further filtered by length to remove low-quality viral genomes (QC1). Finally, all filtered viral species were pooled across cohorts and deduplicated again, yielding the initial vOTU set. Sample reads were mapped back onto this vOTU set in order to obtain viral species abundances per sample. iPhop was run on the non-human-infecting vOTU sequences to obtain prokaryotic host predictions, with contradictory predictions resolved by descending to the last common ancestor (LCA). Next, proteins encoded within each viral species were compared to proteins from all other viral species to derive viral gene families (VOGs) for functional annotation. Functional annotations were used to further filter the viral species (by requiring structural proteins) into a high-confidence subset (QC2) used for all subsequent association analyses. The all-against-all viral protein comparison was also used to generate a global viral hierarchical clustering, in order to derive viral families (VFCs). Viral protein sequences were embedded using the ProtT5 protein language model to obtain weighted mean protein embeddings for each vOTU. The viral lifestyle for each vOTU was determined by passing the vOTU embeddings through a linear classifier. Lifestyles were also determined at the VFC-level by mean pooling vOTU embeddings by VFC affiliation and running the same linear classifier.


preprocessing
read QC
MIDIA
assembly
MIDIA
deduplication
MIDIA
geNomad
MIDIA
vOTU QC1
MIDIA
read QC
DIPP
assembly
DIPP
deduplication
DIPP
geNomad
DIPP
vOTU QC1
DIPP
read QC
COPSAC
assembly
COPSAC
deduplication
COPSAC
geNomad
COPSAC
vOTU QC1
COPSAC
pool vOTUs
all cohorts
functional annotation
PHROGs
functions
HHblits
PHROGs
MSAs
VOGs
VOGs
alignment
head + tail
vOTU QC2
lifestyle prediction
lifestyle model
VFC pools
VFC pooling
embeddings
lifestyle model
pooled embed.
mean pooling
embeddings
pLM embedding
protein pool
cut tree
VFC threshold
APS tree
distances
vOTU distances
alignment
viral taxonomy
VFCs
alignment
protein pool
extract proteins
vOTUs
deduplication
vOTU pool
host prediction
host range
vOTU LCA
LCA consensus
host predictions
iPhop
vOTUs
abundance
map reads
vOTUs
counts
BAM files
vOTU metadata
- Enhanced QCs
- VFCs
- Bacterial host ranges
- Viral lifestyle
- VFC lifestyle
VFC table
OTU table
host counts
OTU table
OTU table
counts

Supplementary Figure 3. Effect of prevalence filtering on the number of taxa and total abundance.

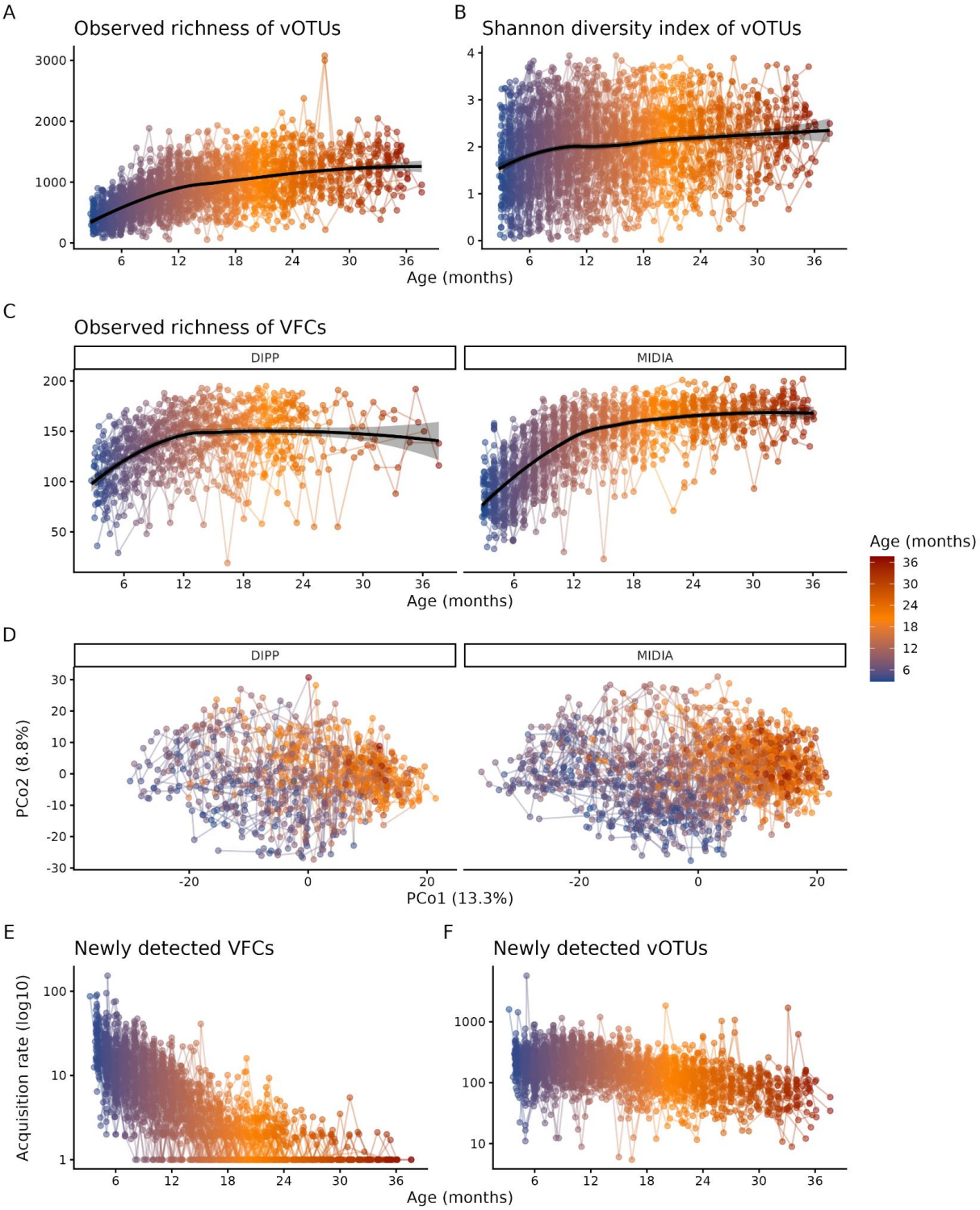
Filtering thresholds based on prevalence (across all samples) and percentage retained reads for the remaining VFCs, plotted at A) the viral family level and B) the viral host genus level. The red line represents the percentage of reads kept and the blue line represents the percentage of VFCs or viral hosts kept at each threshold. The selected filtering is demonstrated with a vertical dashed line (30% for viral families and 10% for viral hosts).



Supplementary Figure 4. Alpha and beta diversity.

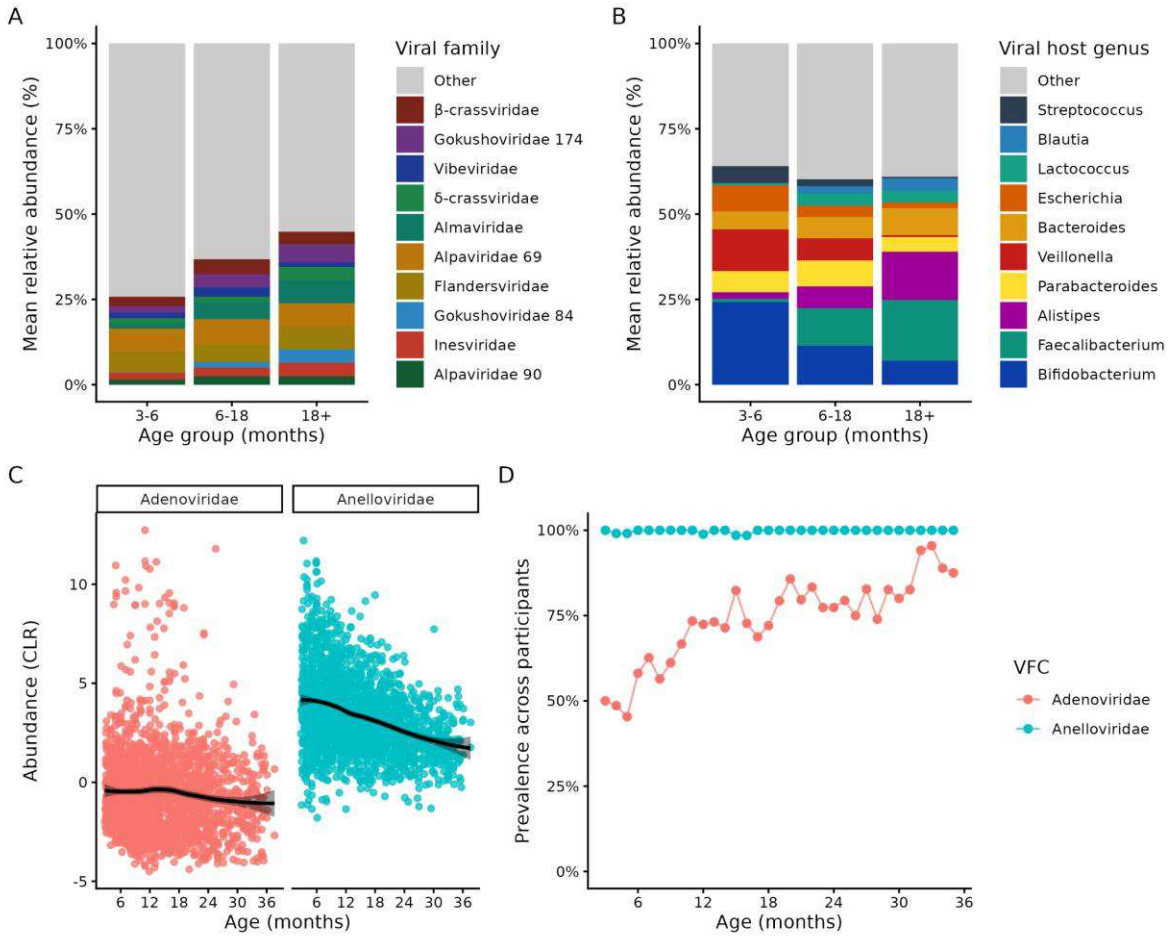
A) Observed richness,
B) Shannon diversity index of viral species (vOTUs) from age 3-36 months.
C) Observed richness of VFCs from age 3-36 months faceted by cohort. The black line represents the LOESS locally smoothed regression fit with 95% confidence intervals (±1.96 standard errors).
D) Principal coordinate analysis of Aitchison distances between samples quantified at the VFC level, faceted by cohort.
E) Newly detected VFCs over time (acquisition rate).
F) Observed richness step size of VFCs per consecutive samples within a child.

The points represent individual samples, and the lightly colored lines connect the samples of individual participants. The points and lines are colored by the age at sampling.


A
Observed richness of vOTUs
B
Shannon diversity index of vOTUs
Age (months)
C
Observed richness of VFCs
DIPP
MIDIA
Age (months)
Age (months)
D
DIPP
MIDIA
PCo2 (8.8%)
PCo1 (13.3%)
E
Newly detected VFCs
F
Newly detected vOTUs
Acquisition rate (log10)
Age (months)

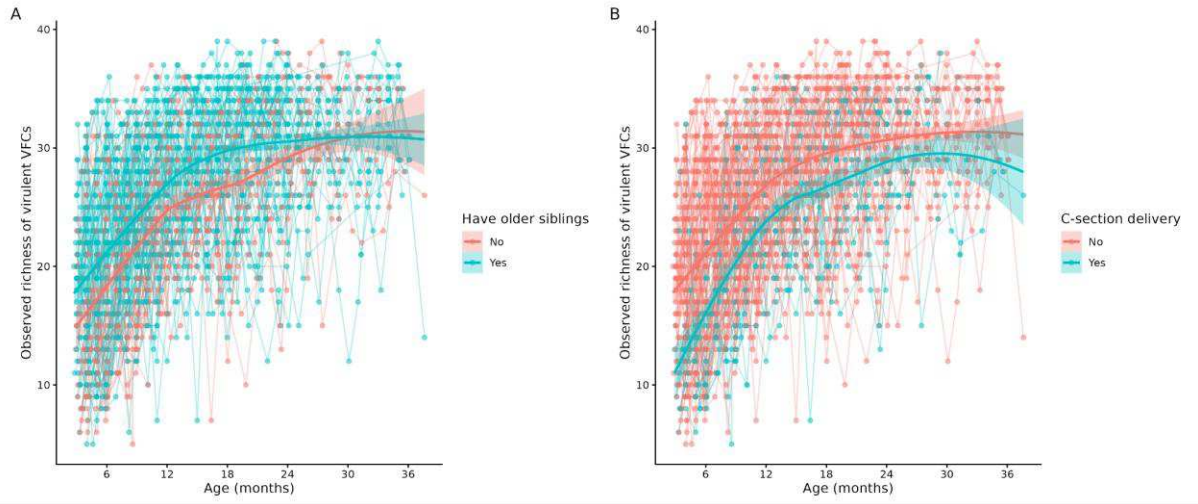
Supplementary Figure 5. Viral phage composition by age groups and the eukaryotic virome.

Stacked barchart of the average relative abundance of A) viral families and B) predicted viral host genera. The columns represent age groups of the samples.
C) Abundance (centered log ratio (CLR) (+1 pseudocount) transformed abundance) of Adenoviridae and Anelloviridae.
D) The prevalence (detected by read mapping) across samples at each age. The prevalence is only plotted if there are more than 10 samples at the given timepoint.



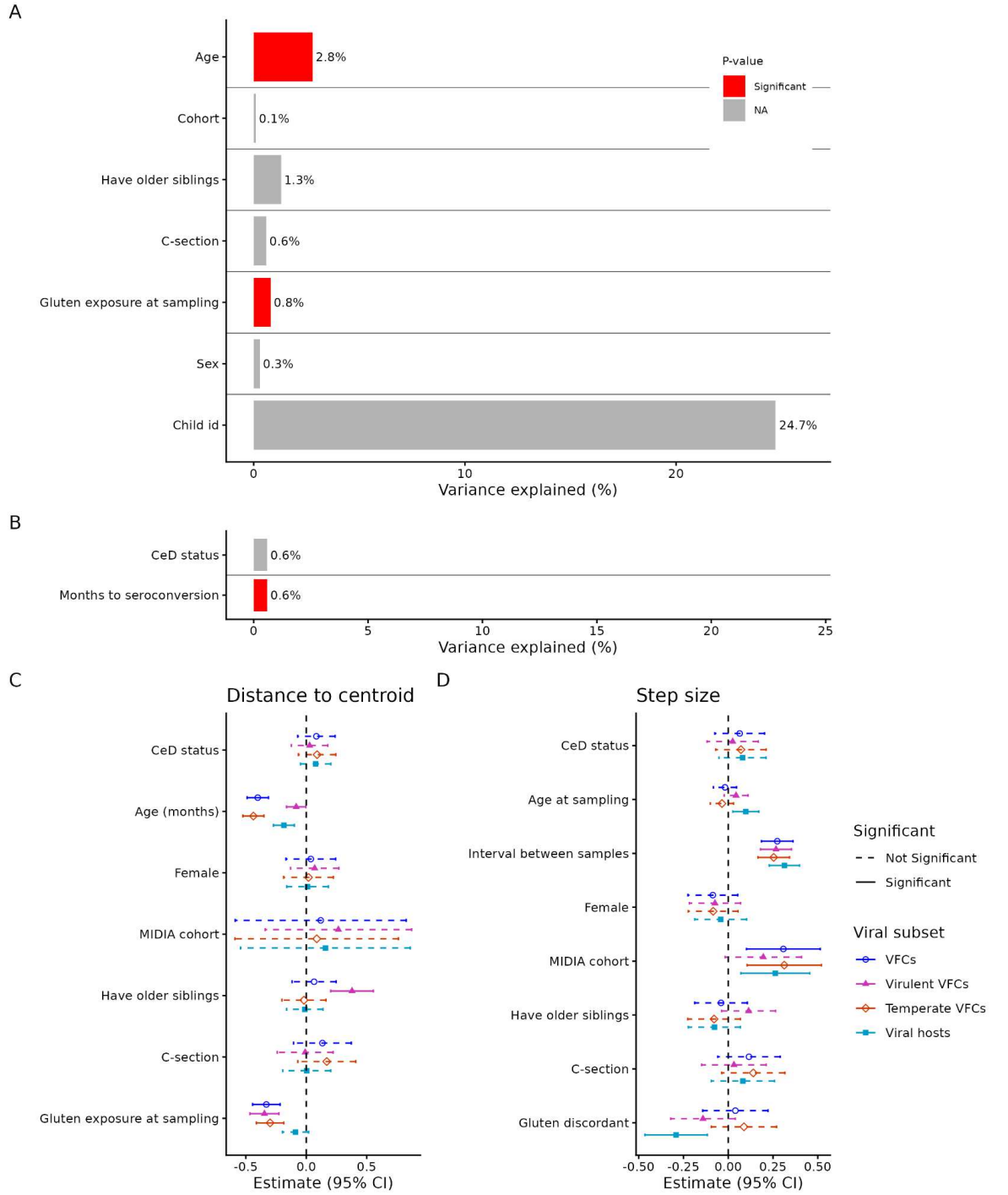
Supplementary Figure 6. Predicted richness of virulent viral families over age.

Split by A) whether the participant has older siblings or B) whether the participant was born by C-section. The points represent the observed richness for each sample. The thick lines represent the GAM predicted fit with 95% confidence intervals (±1.96 standard errors), adjusted for covariates and longitudinal design.



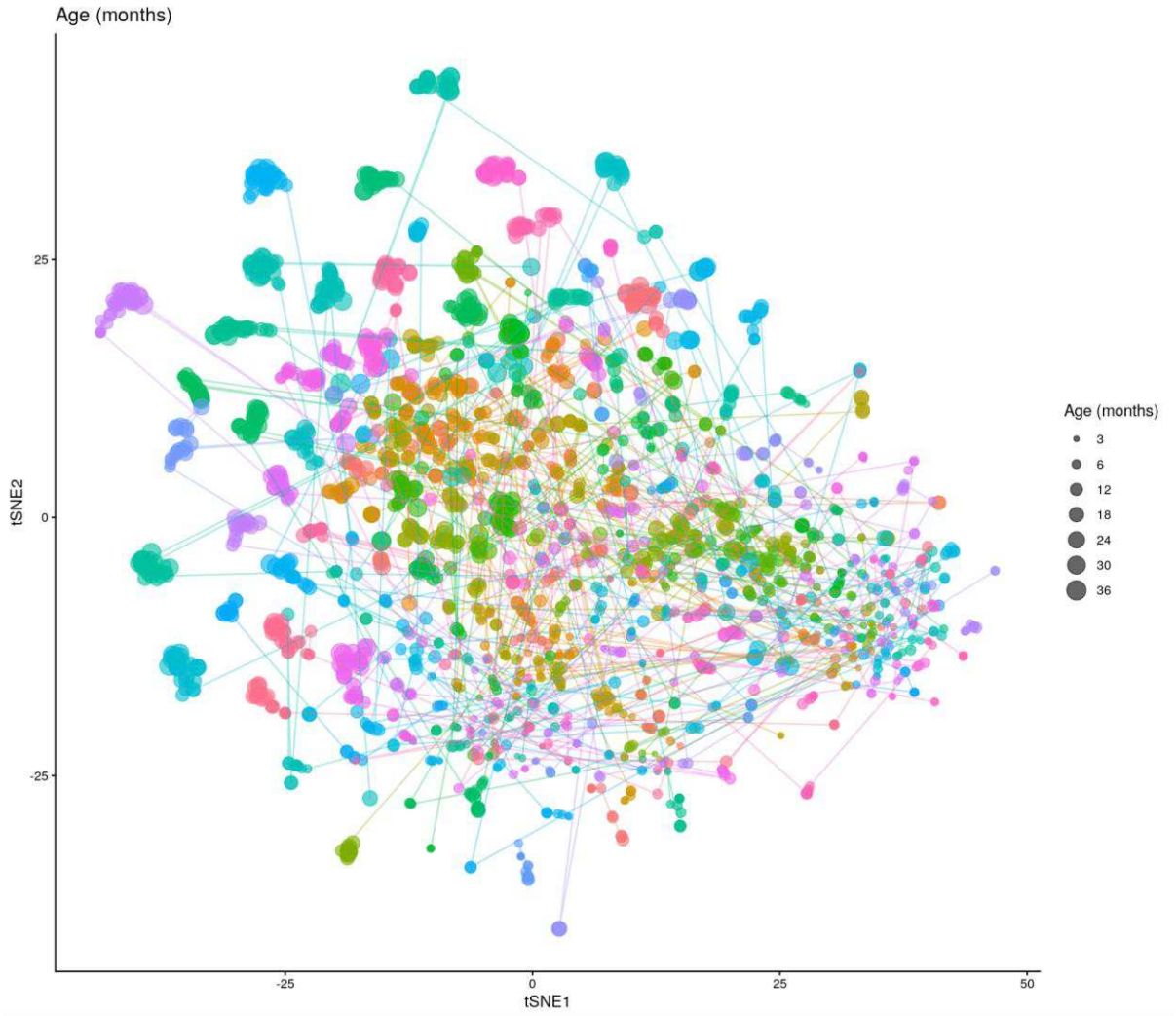
Supplementary Figure 7. Relationship between several factors and the beta diversity as tested by a PERMANOVA, and analyses of distance to centroid and step size.

A) and B) The variance explained by several factors on the beta diversity (Aitchison distance) of all VFCs calculated by a PERMANOVA. The color of the bars indicate statistical significance (red = Significant (p<0.05), grey = unable to test for significance). The PERMANOVA was stratified by participant ID, so it was only possible to test the significance of variables that varied within a participant over time. C) Forest plot of the linear mixed model testing the distance to centroid across all samples per child. D) Forest plot of the linear mixed model testing the step size per consecutive sample per child.



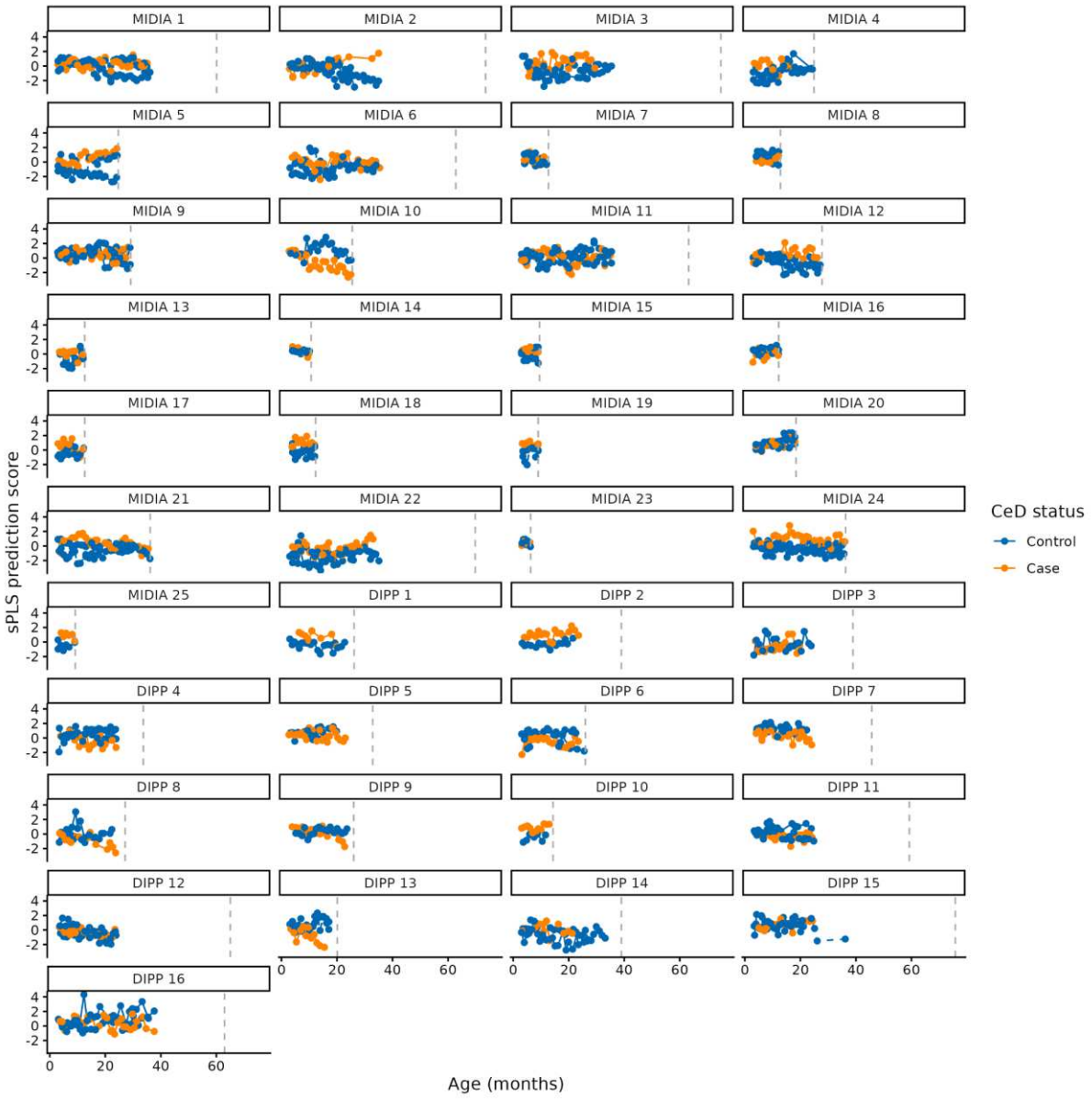
Supplementary Figure 8. t-distributed stochastic neighbor embedding (t-SNE) plot of all VFCs.

Each point represents a sample, colored and connected by participant ID. The size of each point reflects the participant's age at sampling. The t-SNE was performed on CLR-transformed VFC abundances after dimensionality reduction by principal component analysis (PCA), using the first 100 principal components as input.



Supplementary Figure 9. sPLS prediction score over age.

The sPLS prediction score for virulent VFCs is plotted along the Y-axis against the age along the X-axis. Vertical dashed lines denote the time of seroconversion for CeD cases (when the last blood sample was negative for tTGA). Each panel contains the samples from the separate matching groups, and the lines follow each control and case, which are colored blue or orange, respectively.



References

1. Liang, G. & Bushman, F. D. The human virome: assembly, composition and host interactions. *Nature Reviews Microbiology* **19**, 514–527 (2021).

2. Breitbart, M. *et al.* Viral diversity and dynamics in an infant gut. *Res Microbiol* **159**, 367–373 (2008).

3. Virgin, H. W. The virome in mammalian physiology and disease. *Cell* **157**, 142–150 (2014).

4. Champagne-Jorgensen, K., Luong, T., Darby, T. & Roach, D. R. Immunogenicity of bacteriophages. *Trends Microbiol* **31**, 1058–1071 (2023).

5. Leal Rodríguez, C. *et al.* The infant gut virome is associated with preschool asthma risk independently of bacteria. *Nat. Med.* **30**, 138–148 (2024).

6. Liang, G., Gao, H. & Bushman, F. D. The pediatric virome in health and disease. *Cell Host Microbe* **30**, 639–649 (2022).

7. Tapia, G. *et al.* Parechovirus Infection in Early Childhood and Association With Subsequent Celiac Disease. *Am J Gastroenterol* **116**, 788–795 (2021).

8. Boulund, U. *et al.* Infant gut anellovirus composition associates with early life factors and childhood atopic disease. *Nat. Commun* 17, 80 (2026)

9. Koonin, E. V., Dolja, V. V. & Krupovic, M. The healthy human virome: from virus-host symbiosis to disease. *Curr Opin Virol* **47**, 86–94 (2021).

10. Sciurti, M. *et al.* Genetic susceptibilty and celiac disease: what role do HLA haplotypes play? *Acta Biomed* **89**, 17–21 (2018).

11. King, J. A. *et al.* Incidence of Celiac Disease Is Increasing Over Time: A Systematic Review and Meta-analysis. *Am J Gastroenterol* **115**, 507–525 (2020).

12. Gatti, S., Rubio-Tapia, A., Makharia, G. & Catassi, C. Patient and Community Health Global Burden in a World With More Celiac Disease. *Gastroenterology* **167**, 23–33 (2024).

13. Kårhus, L. L., Thuesen, B. H., Skaaby, T., Rumessen, J. J. & Linneberg, A. The distribution of HLA DQ2 and DQ8 haplotypes and their association with health indicators in a general Danish population. *United European Gastroenterol J* **6**, 866–878 (2018).

14. Stahl, M. *et al.* Incidence of Pediatric Celiac Disease Varies by Region. *Am J Gastroenterol* **118**, 539–545 (2023).

15. Af Segerstad, E. M. H. *et al.* Associations of early life dietary diversity and healthy eating with celiac disease: A prospective cohort study. *Clin Nutr* **52**, 171–178 (2025).

16. Taylor, A. K., Lebwohl, B., Snyder, C. L. & Green, P. H. R. Celiac Disease. in *GeneReviews® [Internet]* (University of Washington, Seattle, 2025).

17. Oikarinen, M. *et al.* Enterovirus Infections Are Associated With the Development of Celiac Disease in a Birth Cohort Study. *Front Immunol* **11**, 604529 (2020).

18. Kahrs, C. R. *et al.* Enterovirus as trigger of coeliac disease: nested case-control study within prospective birth cohort. *BMJ* **364**, l231 (2019).

19. Leonard, M. M. *et al.* Microbiome signatures of progression toward celiac disease onset in at-risk children in a longitudinal prospective cohort study. *Proceedings of the National Academy of Sciences* **118**, e2020322118 (2021).

20. Maqsood, R. *et al.* Discordant transmission of bacteria and viruses from mothers to babies at birth. *Microbiome* **7**, 156 (2019).

21. Taboada, B. *et al.* The gut virome of healthy children during the first year of life is diverse and dynamic. *PLoS One* **16**, e0240958 (2021).

22. Zeng, S. *et al.* A metagenomic catalog of the early-life human gut virome. *Nat. Commun.* **15**, 1864 (2024).

23. Reyes, A. *et al.* Gut DNA viromes of Malawian twins discordant for severe acute malnutrition. *Proc Natl Acad Sci U S A* **112**, 11941–11946 (2015).

24. Lim, E. S. *et al.* Early life dynamics of the human gut virome and bacterial microbiome in infants. *Nat. Med.* **21**, 1228–1234 (2015).

25. Walters, W. A. *et al.* Longitudinal comparison of the developing gut virome in infants and their mothers. *Cell Host Microbe* **31**, 187–198.e3 (2023).

26. McCann, A. *et al.* Viromes of one year old infants reveal the impact of birth mode on microbiome diversity. *PeerJ* **6**, e4694 (2018).

27. Garmaeva, S. *et al.* Transmission and dynamics of mother-infant gut viruses during pregnancy and early life. *Nat Commun* **15**, 1945 (2024).

28. Liang, G. *et al.* The stepwise assembly of the neonatal virome is modulated by breastfeeding. *Nature* **581**, 470–474 (2020).

29. Donald, K. & Finlay, B. B. Early-life interactions between the microbiota and immune system: impact on immune system development and atopic disease. *Nat Rev Immunol* **23**, 735–748 (2023).

30. Shah, S. A. *et al.* Expanding known viral diversity in the healthy infant gut. *Nat. Microbiol.* **8**, 986–998 (2023).

31. Kupila, A. *et al.* Feasibility of genetic and immunological prediction of type I diabetes in a population-based birth cohort. *Diabetologia* **44**, 290–297 (2001).

32. Stene, L. C. *et al.* Islet autoantibody development during follow-up of high-risk children from the general Norwegian population from three months of age: design and early results from the MIDIA study. *J Autoimmun* **29**, 44–51 (2007).

33. Camargo, A. P. *et al.* Identification of mobile genetic elements with geNomad. *Nature Biotechnology* **42**, 1303–1312 (2023).

34. Roux, S. *et al.* iPHoP: An integrated machine learning framework to maximize host prediction for metagenome-derived viruses of archaea and bacteria. *PLoS Biol* **21**, e3002083 (2023).

35. Yatsunenko, T. *et al.* Human gut microbiome viewed across age and geography. *Nature* **486**, 222–227 (2012).

36. Roswall, J. *et al.* Developmental trajectory of the healthy human gut microbiota during the first 5 years of life. *Cell Host Microbe* **29**, 765–776.e3 (2021).

37. Stokholm, J. *et al.* Maturation of the gut microbiome and risk of asthma in childhood. *Nat. Commun.* **9**, 141 (2018).

38. Shamash, M. & Maurice, C. F. The infant gut virome assembles rapidly and predictably over the first three years of life. *bioRxiv* (2026) doi:10.64898/2026.02.17.705938.

39. Christensen, E. D. *et al.* The developing airway and gut microbiota in early life is influenced by age of older siblings. *Microbiome* **10**, 106 (2022).

40. Laursen, M. F. *et al.* Having older siblings is associated with gut microbiota development during early childhood. *BMC Microbiol* **15**, 154 (2015).

41. Stokholm, J. *et al.* Delivery mode and gut microbial changes correlate with an increased risk of childhood asthma. *Sci. Transl. Med.* **12**, eaax9929 (2020).

42. Prendergast, P. J. *et al.* Comprehensive cross-cohort analysis reveals global gut microbiome signatures of celiac disease. *Communications Medicine* **6**, 392 (2026).

43. Girdhar, K. *et al.* Dynamics of the gut microbiome, IgA response, and plasma metabolome in the development of pediatric celiac disease. *Microbiome* **11**, 9 (2023).

44. Zafeiropoulou, K. *et al.* Alterations in Intestinal Microbiota of Children With Celiac Disease at the Time of Diagnosis and on a Gluten-free Diet. *Gastroenterology* **159**, 2039–2051.e20 (2020).

45. Lindfors, K. *et al.* Metagenomics of the faecal virome indicate a cumulative effect of enterovirus and gluten amount on the risk of coeliac disease autoimmunity in genetically at risk children: the TEDDY study. *Gut* **69**, 1416–1422 (2020).

46. Lin, J. *et al.* Vipie: web pipeline for parallel characterization of viral populations from multiple NGS samples. *BMC Genomics* **18**, 378 (2017).

47. El Mouzan, M. *et al.* Viral dysbiosis in children with new-onset celiac disease. *PLoS One* **17**, e0262108 (2022).

48. Andeweg, S. P., Keşmir, C. & Dutilh, B. E. Quantifying the Impact of Human Leukocyte Antigen on the Human Gut Microbiota. *mSphere* 6 (2021)

49. Olivares, M. *et al.* The HLA-DQ2 genotype selects for early intestinal microbiota composition in infants at high risk of developing coeliac disease. *Gut* **64**, 406–417 (2015).

50. Boulund, U. *et al.* Gut microbiome associations with host genotype vary across ethnicities and potentially influence cardiometabolic traits. *Cell Host Microbe* **30**, 1464–1480.e6 (2022).

51. Rasmussen, T., Witsø, E., Tapia, G., Stene, L. C. & Rønningen, K. S. Self-reported lower respiratory tract infections and development of islet autoimmunity in children with the type 1 diabetes high-risk HLA genotype: the MIDIA study. *Diabetes Metab Res Rev* **27**, 834–837 (2011).

52. Skrivarhaug, T. *et al.* Incidence of type 1 diabetes in Norway among children aged 0-14 years between 1989 and 2012: has the incidence stopped rising? Results from the Norwegian Childhood Diabetes Registry. *Diabetologia* **57**, 57–62 (2014).

53. Cinek, O. *et al.* Longitudinal observation of enterovirus and adenovirus in stool samples from Norwegian infants with the highest genetic risk of type 1 diabetes. *J Clin Virol* **35**, 33–40 (2006).

54. Salminen, K. K. *et al.* Isolation of enterovirus strains from children with preclinical Type 1 diabetes. *Diabet Med* **21**, 156–164 (2004).

55. Cinek, O. *et al.* Eukaryotic viruses in the fecal virome at the onset of type 1 diabetes: A study from four geographically distant African and Asian countries. *Pediatr. Diabetes* **22**, 558–566 (2021).

56. Lopez-Labrador, F. X. *et al.* Multicenter benchmarking of short and long read wet lab protocols for clinical viral metagenomics. *J Clin Virol* **173**, 105695 (2024).

57. Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**, e2584 (2016).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- HEDIMEDviromesSupplementaryTables.xlsx