Supplemental Information

The Dynamics of the Human Infant Gut Microbiome in Development and in Progression towards Type 1 Diabetes

Aleksandar D. Kostic, Dirk Gevers, Heli Siljander, Tommi Vatanen, Tuulia Hyötyläinen, Anu-Maaria Hämäläinen, Aleksandr Peet, Vallo Tillmann, Päivi Pöho, Ismo Mattila, Harri Lähdesmäki, Eric A. Franzosa, Outi Vaarala, Marcus de Goffau, Hermie Harmsen, Jorma Ilonen, Suvi M. Virtanen, Clary B. Clish, Matej Orešič, Curtis Huttenhower, Mikael Knip* on behalf of the DIABIMMUNE Study Group, and Ramnik J. Xavier*

*shared senior authorship

Supplemental Information Inventory

Supplemental Author Information

The DIABIMMUNE Study Group comprises the following investigators: Mikael Knip,PI (Children's Hospital, University of Helsinki), Katriina Koski, Coordinator (Institute of Clinical Medicine, University of Helsinki), Matti Koski (IT Manager, Institute of Clinical Medicine, University of Helsinki), Taina Härkönen (Children's Hospital, University of Helsinki), Samppa Ryhänen (Children's Hospital, University of Helsinki), Heli Siljander (Children's Hospital, University of Helsinki), Anu-Maaria Hämäläinen (Jorvi Hospital, Helsinki University Central Hospital), Anne Ormisson (Children's Clinic, Tartu University Hospital), Aleksandr Peet (Department of Pediatrics, Tartu University Hospital), Vallo Tillmann (Department of Pediatrics, Tartu University Hospital), Valentina Ulich (Ministry of Health and Social Development, Karelian Republic of the Russian Federation), Elena Kuzmicheva (Ministry of Health and Social Development, Karelian Republic of the Russian Federation), Sergei Mokurov (Ministry of Health and Social Development, Karelian Republic of the Russian Federation), Svetlana Markova (Children's Republic Hospital, Karelian Republic of the Russian Federation), Svetlana Pylova (Children's Republic Hospital, Karelian Republic of the Russian Federation), Marina Isakova (Perinatal Center, Karelian Republic of the Russian Federation), Elena Shakurova (Perinatal Center, Karelian Republic of the Russian Federation), Vladimir Petrov (Maternity Hospital N° 1, Petrozavodsk), Natalya V. Dorshakova (Petrozavodsk State University), Tatyana Karapetyan (Petrozavodsk State University), Tatyana Varlamova (Petrozavodsk State University), Jorma Ilonen (Immunogenetics Laboratory, University of Turku and Department of Clinical Microbiology, University of Eastern Finland, Kuopio), Minna Kiviniemi (Immunogenetics Laboratory, University of Turku), Kristi Alnek (Department of Immunology, University of Tartu), Helis Janson (Department of Immunology, University of Tartu) Raivo Uibo (Department of Immunology, University of Tartu), Tiit Salum (OÜ Immunotron, Tartu), Erika von Mutius (Children's Hospital, Ludwig Maximilians Universität, Munchen), Juliane Weber (Children's Hospital, Ludwig Maximilians Universität, Munchen), Helena Ahlfors (Turku Centre of Biotechnology, University of Turku and Åbo Akademi), Henna Kallionpää, (Turku Centre of Biotechnology, University of Turku and Åbo Akademi), Essi Laajala (Turku Centre of Biotechnology, University of Turku and Åbo Akademi), Riitta Lahesmaa (Turku Centre of Biotechnology, University of Turku and Åbo Akademi), Harri Lähdesmäki (Turku Centre of Biotechnology, University of Turku and Åbo Akademi), Robert Moulder (Turku Centre of Biotechnology, University of Turku and Åbo Akademi), Viveka Öling (Turku Centre of Biotechnology, University of Turku and Åbo Akademi), Janne Nieminen (Department of Vaccination and Immune Protection, National Institute for Health and Welfare, Helsinki), Terhi Ruohtula (Department of Vaccination and Immune Protection, National Institute for Health and Welfare, Helsinki), Outi Vaarala (Department of Vaccination and Immune Protection, National Institute for Health and Welfare, Helsinki), Hanna Honkanen (Department of Virology, University of Tampere), Heikki Hyöty (Department of Virology, University of Tampere and Tampere University Hospital), Anita Kondrashova (Department of Virology, University of Tampere), Sami Oikarinen (Department of Virology, University of Tampere), Hermie J.M. Harmsen (University Medical Center Groningen), Marcus C. De Goffau (University Medical Center Groningen), Gjal Welling (University Medical Center Groningen), Kirsi Alahuhta (Department for Welfare and Health Promotion, National Institute for Health and Welfare, Helsinki), Tuuli Korhonen (Department of Life Style and Participation, National Institute for Health and Welfare, Helsinki), Suvi M. Virtanen (Department of Life Style and Participation, National Institute for Health and Welfare, Helsinki and School of Health Sciences, University of Tampere, Tampere), Taina Öhman (Department of Life Style and Participation, National Institute for Health and Welfare, Helsinki).

Supplemental Experimental Procedures

Study Cohort

The children were recruited to the study cohort from Espoo, Finland (*n* = 27) and Tartu, Estonia (*n* = 6) between September 2008 and August 2010 as part of the international DIABIMMUNE Study (http://www.diabimmune.org/). Inclusion criteria included positive cord blood testing for HLA DR-DQ alleles conferring risk of T1D. The participating children were monitored prospectively for infections, use of drugs, antibiotics in particular, and other life events. Data on breastfeeding and introduction of complementary foods was registered continuously. Serum samples were collected from all infants during visits to the clinic at the following time-points: 0 (cord blood), 3, 6, 12, 18, 24, and 36 months. The following 4 diabetes-associated autoantibodies were analyzed from each serum sample with specific radiobinding assays: insulin autoantibodies (IAA), glutamic acid decarboxylase antibodies (GADA), islet antigen-2 antibodies (IA-2A), and zinc transporter 8 antibodies (ZnT8A) as described (Knip et al., 2010). Islet cell antibodies (ICA) were analyzed with immunofluoresence in those subjects who tested positive for at least one of the biochemical autoantibodies. The cut-off values were based on the 99th percentile in non-diabetic children and were 2.80 relative units (RU) for IAA, 5.36 RU for GADA, 0.78 RU for IA-2A and 0.61 RU for ZnT8A, The detection limit in the ICA assay was 2.5 Juvenile Diabetes Foundation units (JDFU).
Subject metadata is available in **Table S1**.

Sequencing and Analysis of the 16S Gene

The 16S gene was sequenced on the Illumina MiSeq platform targeting the V4 variable region. Detailed protocols used for 16S amplification and sequencing are as previously described (Caporaso et al., 2012). Briefly, genomic DNA was subjected to 16S amplifications using primers incorporating the Illumina adapters and a sample barcode sequence, allowing directional sequencing covering variable region V4 (Primers: 515F [GTGCCAGCMGCCGCGGTAA] and 806R [GGACTACHVGGGTWTCTAA]). The PCR mix contained 10 µl of diluted template (1:50), 10 µl of HotMasterMix with the HotMaster Taq DNA Polymerase (5 Prime), and 5 µl of primer mix (2 µM of each primer). The cycling conditions were as follows: 94°C for 3 min, then 30 cycles of denaturation at 94°C for 45 sec, annealing at 50°C for 60 sec, extension at 72°C for 5 min, and a final extension at 72°C for 10 min. Amplicons were quantified on the Caliper LabChipGX (PerkinElmer, Waltham, MA), pooled in equimolar concentrations, size selected (375-425 bp) on the Pippin Prep (Sage Sciences, Beverly, MA) to reduce non-specific amplification products from host DNA, and a final library size and quantification was performed on the Agilent Bioanalyzer 2100 DNA1000 chips (Agilent Technologies, Santa Clara, CA). Sequencing was performed on the Illumina MiSeq v2 platform, according to the manufacturer's specifications with spike-in of 5% PhiX DNA, and generating paired-end reads of 175bp in length in each direction. The overlapping paired-end reads

were combined (approximately 97 bp overlap), size selected to reduce non-specific amplification products from host DNA (225 - 275 bp), and further processed in a data curation pipeline implemented in QIIME 1.5.0 as pick_reference_otus.py (Caporaso et al., 2010). In brief, this pipeline picks OTUs using a reference-based method and constructs an OTU table. Taxonomy was assigned using the Greengenes predefined taxonomy map of reference sequence OTUs to taxonomy (McDonald et al., 2012). The resulting OTU tables were checked for mislabeling (Knights et al., 2011a) and contamination (Knights et al., 2011b). A mean sequence depth of 65,076/sample was obtained, and samples with less than 3,000 filtered sequences were excluded from analysis. Functional profiles were predicted from these16S-based microbial compositions using the PICRUSt algorithm (Langille et al., 2013).

Sequencing and Analysis of Shotgun Metagenomics

Metagenomic data production and processing were performed as described previously (Consortium, 2012). In brief, library construction was performed (protocol available: http://hmpdacc.org/tools_protocols/tools_protocols.php) and libraries were sequenced on the Illumina HiSeq 2500 platform, targeting ~2.5Gb of sequence per sample with 101bp, paired-end reads. Contaminating human sequences were filtered-out using BMTagger (http://casbioinfo.cas.unt.edu/sop/mediawiki/index.php/Bmtagger), as were any reads less than 60 nt in length and low-quality reads (BWA-42 quality score <2). Profiling of metabolic pathways was performed by USEARCH version 5.2.32 alignment (Edgar, 2010) to the KEGG orthology database (version April 1, 2013). Best-hits with E<1 were then input into HUMAnN (Abubucker et al., 2012), producing abundance tables for KEGG pathways and modules. Species abundances were calculated with MetaPhlAn 1.7.7 (Segata et al., 2012), following Bowtie 2-2.1.0 alignment (Langmead and Salzberg, 2012) to the MetaPhlAn 1.0 unique marker database (http://huttenhower.sph.harvard.edu/metaphlan).

Strain-level Analysis of Metagenomics Data

Strains were identified using MetaPhlAn strain-level marker genes. Analysis was restricted to only highly abundant species to rule out that the absence of a marker could be attributed to insufficient sequence coverage. Our analysis was restricted to, on a per-individual level, species that had an average MetaPhlAn coverage of at least 1x across all time-points. Marker profiles were converted from abundances to a binary presence vs. absence measure and only markers with variable presence/absence across samples were used in analysis. The fraction of all intra-subject and inter-subject discordant marker pairs were calculated.

Principle Coordinates Analysis

PCoA plots were generated using t-Distributed Stochastic Neighbor Embedding (t-SNE, R implementation was used) (der Maaten and Hinton, 2008) where the free parameter, perplexity, was selected by generating mappings with perplexity values 10, 20, 30, 40 and 50. With 16S data, the algorithm consistently converged to solutions with lowest cost

function with perplexity equal to 20, which value was used for the publication figure. For metabolomics data, perplexity equals 40 was selected using a similar procedure. All metabolomics measurements, including unidentified compounds were used in generating the PCoA plot for metabolomics.

Global lipid and metabolite profiling from serum samples

Two analytical platforms for metabolomics were applied to all samples from the estimation cohort: (1) global lipidomics platform based on Ultra Performance Liquid Chromatography coupled to Mass Spectrometry (UPLC-MS) which covers molecular lipids such asphospholipids, sphingolipids, and neutral lipids (Nygren et al., 2011); (2) platform for combined global profiling of small polar metabolites and quantitative target analysis of selected metabolites based on comprehensive two-dimensional gas chromatography coupled to time-of-flight mass spectrometry (GC×GC-TOFMS) which covers small molecules such as amino acids, free fatty acids, keto-acids, various other organic acids, sterols, and sugars (Castillo et al., 2011). Raw UPLC-MS and GC×GC-TOFMS data were processed with MZmine 2 (Pluskal et al., 2010) and Guineu (Castillo et al., 2011) software, respectively. The final dataset from each platform consisted of a list of metabolite peaks (identified or unidentified) and their levels, calculated using the platform-specific methods, across all samples. Lipids have been named according to the following abbreviations: Cer: ceramide; ChoE: cholesteryl ester; lysoPC: lysophosphatidylcholine; PA: phosphatidic acid; PG: phosphatidylglycerol; PC: phosphatidylcholine; PS: phosphatidylserine; SM: sphingomyelin; TG: triglyceride. The total number of carbons and double bonds for each lipid is indicated.

Metabolomic analysis from stool samples

Stool samples (weight range 50.5-167.8 mg) were homogenized in 4 µL of water per milligram stool sample weight using a bead mill (TissueLyser II; Qiagen) and the aqueous homogenates were aliquoted for metabolite profiling analyses. Four separate liquid chromatography tandem mass spectrometry (LC-MS) methods were used to measure polar metabolites and lipids in each sample. Methods 1, 2 and 3 below were conducted using two LC-MS systems comprised of Nexera X2 U-HPLC systems (Shimadzu Scientific Instruments; Marlborough, MA) and Q Exactive hybrid quadrupole orbitrap mass spectrometers (Thermo Fisher Scientific; Waltham, MA) and method 4 was conducted using a Nexera X2 U-HPLC (Shimadzu Scientific Instruments; Marlborough, MA) coupled to an Exactive Plus orbitrap MS (Thermo Fisher Scientific; Waltham, MA). *Method 1 - positive ion mode MS analyses of polar metabolites*. LC-MS samples were prepared from stool homogenates (10 µL) via protein precipitation with the addition of nine volumes of 74.9:24.9:0.2 v/v/v acetonitrile/methanol/formic acid containing stable isotope-labeled internal standards (valine-d8, Isotec; and phenylalanine-d8, Cambridge Isotope Laboratories; Andover, MA). The samples are centrifuged (10 min, 9,000 x g, 4°C), and the supernatants were injected directly onto a 150 x 2 mm Atlantis HILIC column (Waters; Milford, MA). The column was eluted isocratically at a flow rate of 250 µL/min with 5%

mobile phase A (10 mM ammonium formate and 0.1% formic acid in water) for 1 minute followed by a linear gradient to 40% mobile phase B (acetonitrile with 0.1% formic acid) over 10 minutes. MS analyses were carried out using electrospray ionization in the positive ion mode using full scan analysis over *m/z* 70-800 at 70,000 resolution and 3 Hz data acquisition rate. Additional MS settings were: ion spray voltage, 3.5 kV; capillary temperature, 350°C; probe heater temperature, 300 °C; sheath gas, 40; auxiliary gas, 15; and S-lens RF level 40. *Method 2 – negative ion mode MS analysis of polar metabolites.* LC-MS samples were prepared from stool homogenates (30 µL) via protein precipitation with the addition of four volumes of 80% methanol containing inosine-^{15}N4, thymine-d4 and glycocholate-d4 internal standards (Cambridge Isotope Laboratories; Andover, MA). The samples were centrifuged (10 min, 9,000 x g, 4°C) and the supernatants were injected directly onto a 150 x 2.0 mm Luna NH2 column (Phenomenex; Torrance, CA). The column was eluted at a flow rate of 400 µL/min with initial conditions of 10% mobile phase A (20 mM ammonium acetate and 20 mM ammonium hydroxide in water) and 90% mobile phase B (10 mM ammonium hydroxide in 75:25 v/v acetonitrile/methanol) followed by a 10 min linear gradient to 100% mobile phase A. MS analyses were carried out using electrospray ionization in the negative ion mode using full scan analysis over *m/z* 60-750 at 70,000 resolution and 3 Hz data acquisition rate. Additional MS settings were: ion spray voltage, -3.0 kV; capillary temperature, 350°C; probe heater temperature, 325 °C; sheath gas, 55; auxiliary gas, 10; and S-lens RF level 40. *Method 3 – negative ion mode analysis of metabolites of intermediate polarity (e.g. bile acids and free fatty acids)*. Stool homogenates (30 µL) were extracted using 90 µL of methanol containing PGE2-d4 as an internal standard (Cayman Chemical Co.; Ann Arbor, MI) and centrifuged (10 min, 9,000 x g, 4°C). The supernatents (10 µL) were injected onto a 150 x 2 mm ACQUITY T3 column (Waters; Milford, MA). The column was eluted isocratically at a flow rate of 400 µL/min with 25% mobile phase A (0.1% formic acid in water) for 1 minute followed by a linear gradient to 100% mobile phase B (acetonitrile with 0.1% formic acid) over 11 minutes. MS analyses were carried out using electrospray ionization in the negative ion mode using full scan analysis over *m/z* 200-550 at 70,000 resolution and 3 Hz data acquisition rate. Additional MS settings were: ion spray voltage, -3.5 kV; capillary temperature, 320°C; probe heater temperature, 300 °C; sheath gas, 45; auxiliary gas, 10; and S-lens RF level 60. *Method 4 - polar and nonpolar lipids.* Lipids were extracted from stool homogenates (10 µL) using 190 µL of isopropanol containing 1-dodecanoyl-2-tridecanoyl-sn-glycero-3-phosphocholine as an internal standard (Avanti Polar Lipids; Alabaster, AL). After centrifugation (10 min, 9,000 x g, ambient temperature), supernatants (10 µL) were injected directly onto a 100 x 2.1 mm ACQUITY BEH C8 column (1.7 µm; Waters; Milford, MA). The column was eluted at a flow rate of 450 µL/min isocratically for 1 minute at 80% mobile phase A (95:5:0.1 vol/vol/vol 10 mM ammonium acetate/methanol/acetic acid), followed by a linear gradient to 80% mobile-phase B (99.9:0.1 vol/vol methanol/acetic acid) over 2 minutes, a linear gradient to 100% mobile phase B over 7 minutes, and then 3 minutes at 100% mobile-phase B. MS analyses were carried out using electrospray ionization in the positive ion mode using full scan analysis over *m/z* 200-1100 at 70,000 resolution and 3 Hz data acquisition rate. Additional MS settings were: ion spray voltage, 3.0 kV; capillary

temperature, 300°C; probe heater temperature, 300 °C; sheath gas, 50; auxiliary gas, 15; and S-lens RF level 60. All raw data were processed using Progenesis CoMet software (version 2.0, NonLinear Dynamics) for feature alignment, non-targeted signal detection, and signal integration. Targeted processing of a subset of known metabolites was conducted using TraceFinder software (version 3.0, Thermo Fisher Scientific; Waltham, MA). Compound identities were confirmed using reference standards and reference samples.

Correlations of Stool and Serum Metabolomics and Lipidomics with the Gut Microbiota

For the serum metabolomics and lipidomics measurements, each sample was paired with the closest stool sample before the serum collection time. Between-subject variance and time dependencies in the data were corrected using a linear random effects model ('lmer' function in the R package 'lme4') with subject and serum collection time as separate random effects. The Spearman correlations were computed using the residuals of the model. **Figure 5B** and **Figure S12** use a linear model which does not account for longitudinal effects. To account for multiple comparisons, *P*-values were adjusted using the Benjamini & Hochberg false discovery rate method for each heatmap separately. Heatmaps include microbial clades, lipids and metabolites that have at least one correlation with a non-adjusted *P*-value < 0.01 for genus-level heatmaps, and a stricter selection threshold was used in order to limit the size of the heatmaps on the OTU-level: *P* < 0.005 for metabolites and OTUs, and *P* < 0.001 for lipids and OTUs. For the heatmap in **Figure 5C**, we show only correlations taxa and metabolites with adjusted *P*-value < 0.01. Stool metabolomics and gut microbiota associations were analyzed using penalized canonical correlation analysis (Witten et al., 2009). Penalty terms were chosen to limit the number of genera and compounds used in each canonical variate. Significance of the resulting components was determined using random permutations for both of the data sets.

Supplemental References

Abubucker, S., Segata, N., Goll, J., Schubert, A.M., Izard, J., Cantarel, B.L., Rodriguez-Mueller, B., Zucker, J., Thiagarajan, M., Henrissat, B., et al. (2012). Metabolic Reconstruction for Metagenomic Data and Its Application to the Human Microbiome. PLoS Comput. Biol. *8*, e1002358.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Peña, A.G., Goodrich, J.K., Gordon, J.I., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. Nat. Methods *7*, 335–336.

Caporaso, J.G., Lauber, C.L., Walters, W.A., Berg-Lyons, D., Huntley, J., Fierer, N., Owens, S.M., Betley, J., Fraser, L., Bauer, M., et al. (2012). Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. ISME J. *6*, 1621–1624.

Castillo, S., Mattila, I., Miettinen, J., Orešič, M., and Hyötyläinen, T. (2011). Data analysis tool for comprehensive two-dimensional gas chromatography/time-of-flight mass spectrometry. Anal. Chem. *83*, 3058–3067.

Consortium, T.H.M.P. (2012). Structure, function and diversity of the healthy human microbiome. Nature *486*, 207–214.

Edgar, R.C. (2010). Search and clustering orders of magnitude faster than BLAST. Bioinformatics *26*, 2460–2461.

Knights, D., Kuczynski, J., Koren, O., Ley, R.E., Field, D., Knight, R., DeSantis, T.Z., and Kelley, S.T. (2011a). Supervised classification of microbiota mitigates mislabeling errors. ISME J. *5*, 570–573.

Knights, D., Kuczynski, J., Charlson, E.S., Zaneveld, J., Mozer, M.C., Collman, R.G., Bushman, F.D., Knight, R., and Kelley, S.T. (2011b). Bayesian community-wide culture-independent microbial source tracking. Nat. Methods *8*, 761–763.

Knip, M., Virtanen, S.M., Seppä, K., Ilonen, J., Savilahti, E., Vaarala, O., Reunanen, A., Teramo, K., Hämäläinen, A.-M., Paronen, J., et al. (2010). Dietary intervention in infancy and later signs of beta-cell autoimmunity. N. Engl. J. Med. *363*, 1900–1908.

Langille, M.G.I., Zaneveld, J., Caporaso, J.G., McDonald, D., Knights, D., Reyes, J.A., Clemente, J.C., Burkepile, D.E., Vega Thurber, R.L., Knight, R., et al. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat. Biotechnol.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. J. Mach. Learn. Res. *9*, 85.

McDonald, D., Price, M.N., Goodrich, J., Nawrocki, E.P., DeSantis, T.Z., Probst, A., Andersen, G.L., Knight, R., and Hugenholtz, P. (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J. *6*, 610–618.

Nygren, H., Seppänen-Laakso, T., Castillo, S., Hyötyläinen, T., and Orešič, M. (2011). Liquid chromatography-mass spectrometry (LC-MS)-based lipidomics for studies of body fluids and tissues. Methods Mol. Biol. *708*, 247–257.

Pluskal, T., Castillo, S., Villar-Briones, A., and Oresic, M. (2010). MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC Bioinformatics *11*, 395.

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., and Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. Nat. Methods *9*, 811–814.

Witten, D.M., Tibshirani, R., and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics *10*, 515–534.

Supplemental Figures

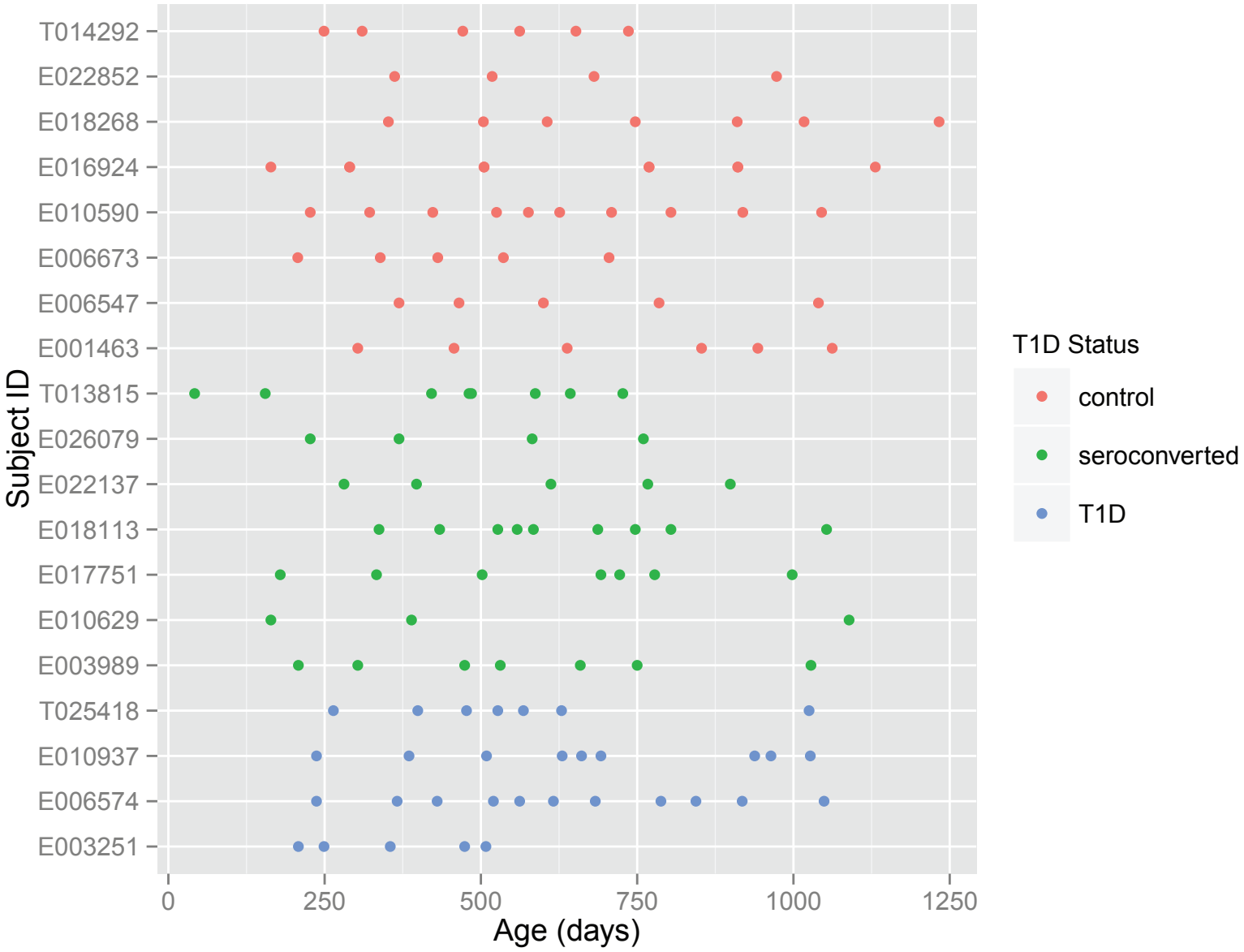

Figure S1, related to Figure 1. Shotgun metagenomic sequencing on a subset of samples. Schematic of the subset of stool samples shown in **Fig. 1A** that have been analyzed by metagenomics. Individuals are represented in rows and each point is a stool sample.

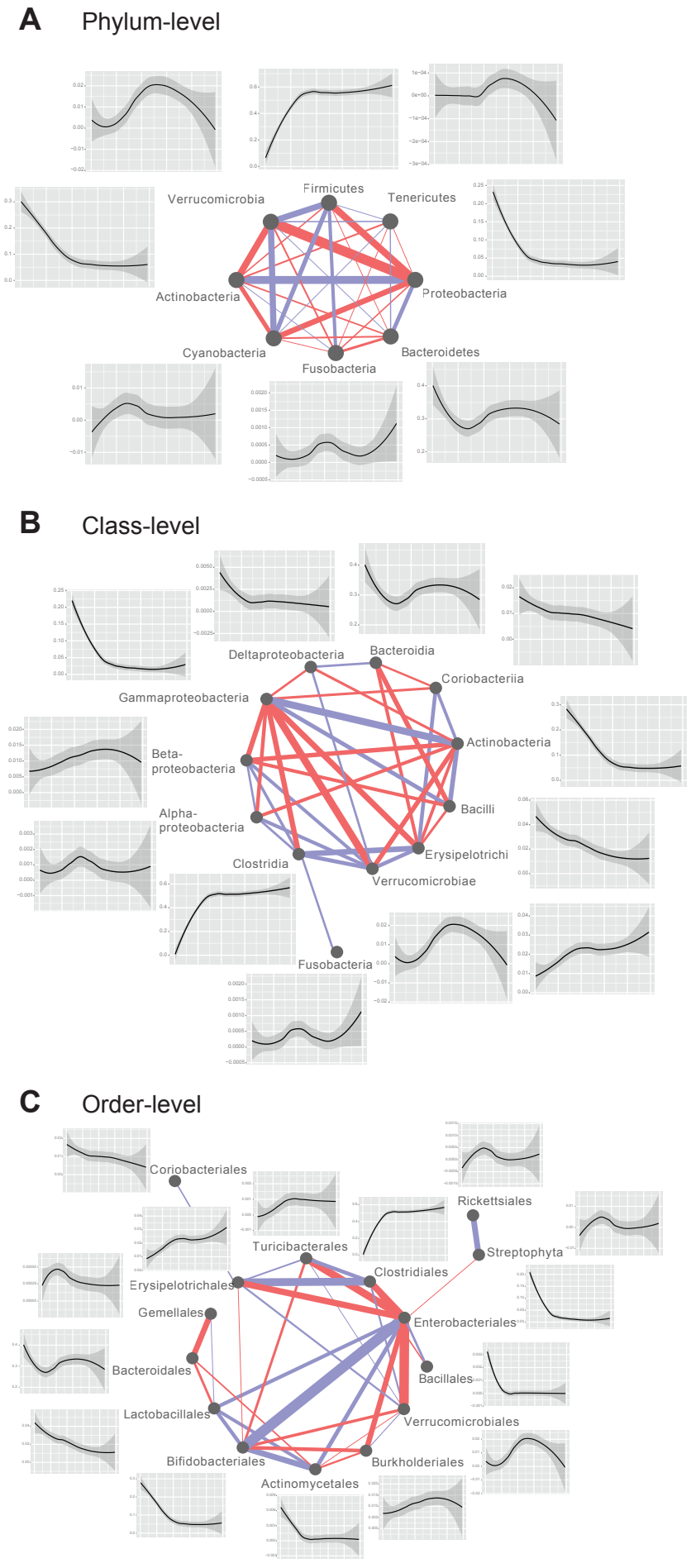

Figure S2, related to Figure 3. Dynamics of the developing gut microbiota. Network diagrams of correlations in trajectory are shown as in **Fig. 1C** for the following phylogenetic levels: (**A**) phylum, (**B**) class, (**C**) order, and (**D**) genus levels.

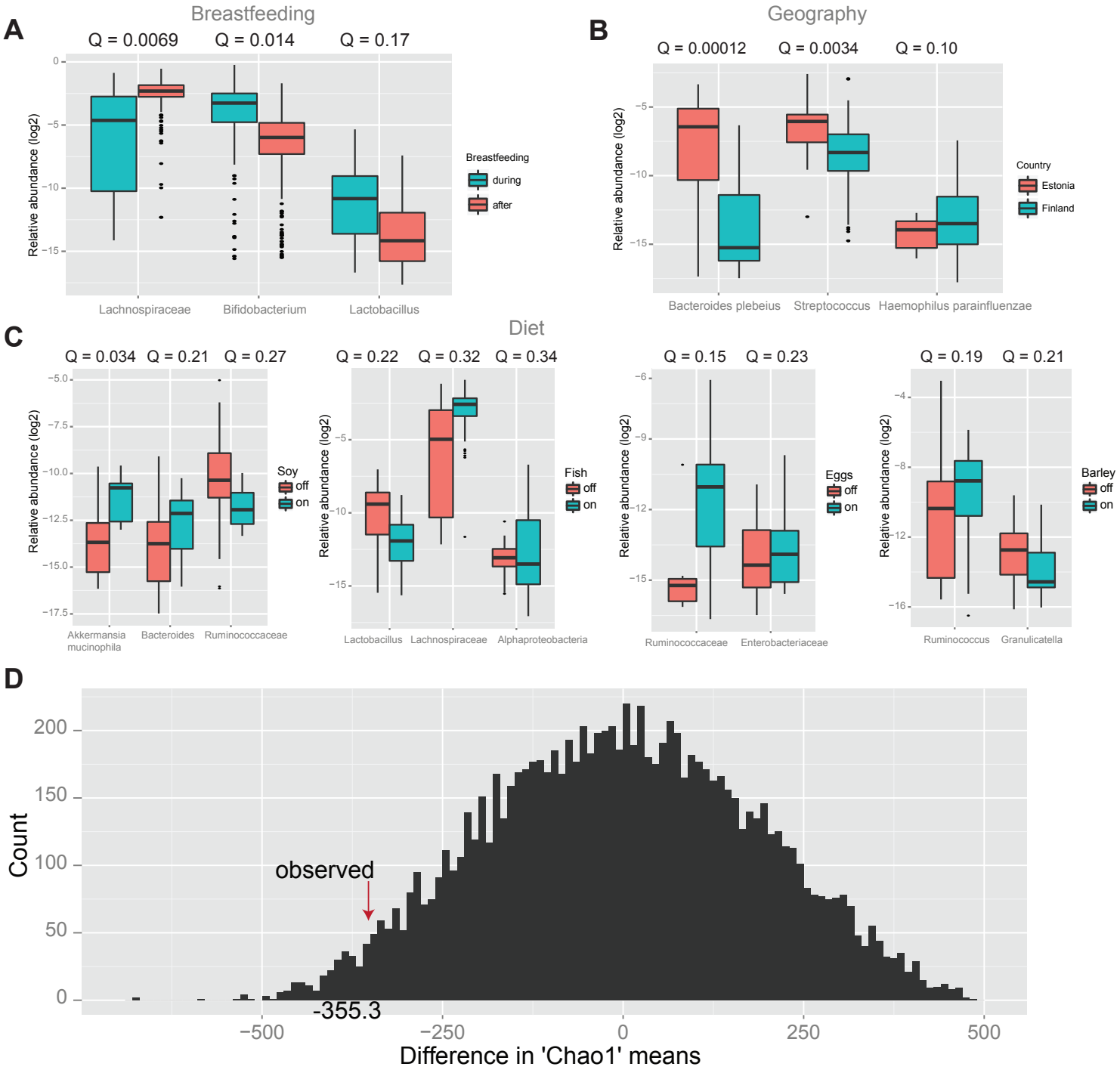

Figure S3, related to Figure 5. Factors that shape the gut microbiota; decreases in alpha-diversity in T1D subjects are significant by permutation test on subject labels. (**A**) Significant community shifts associated with breastfeeding from a series of five successive samples from each individual selected during and after breastfeeding; excludes non-breastfed individuals (n=1). (**B**) Significant differences between Finnish and Estonian individuals based on samples selected during a time-window of 500 to 750 days of age (~6 samples per individual). (**C**) Significant changes in community composition before and after introduction of the indicated food to the diet; a maximum of 8 samples were selected from each individuals around the time of the change in diet. All analyses were performed and Q-values computed with MaAsLin. (**D**) Permutation-based analysis of the significance of the difference in Chao1 means. 10,000 subject-based permutations were performed on all 33 individuals and the difference in the Chao1 (alpha-diversity) mean between the T1D group (n=4) and the control and seroconverted group (n=29) was calculated. The observed differences in Chao1 means, -355.3, is indicated. $P < 0.025$.

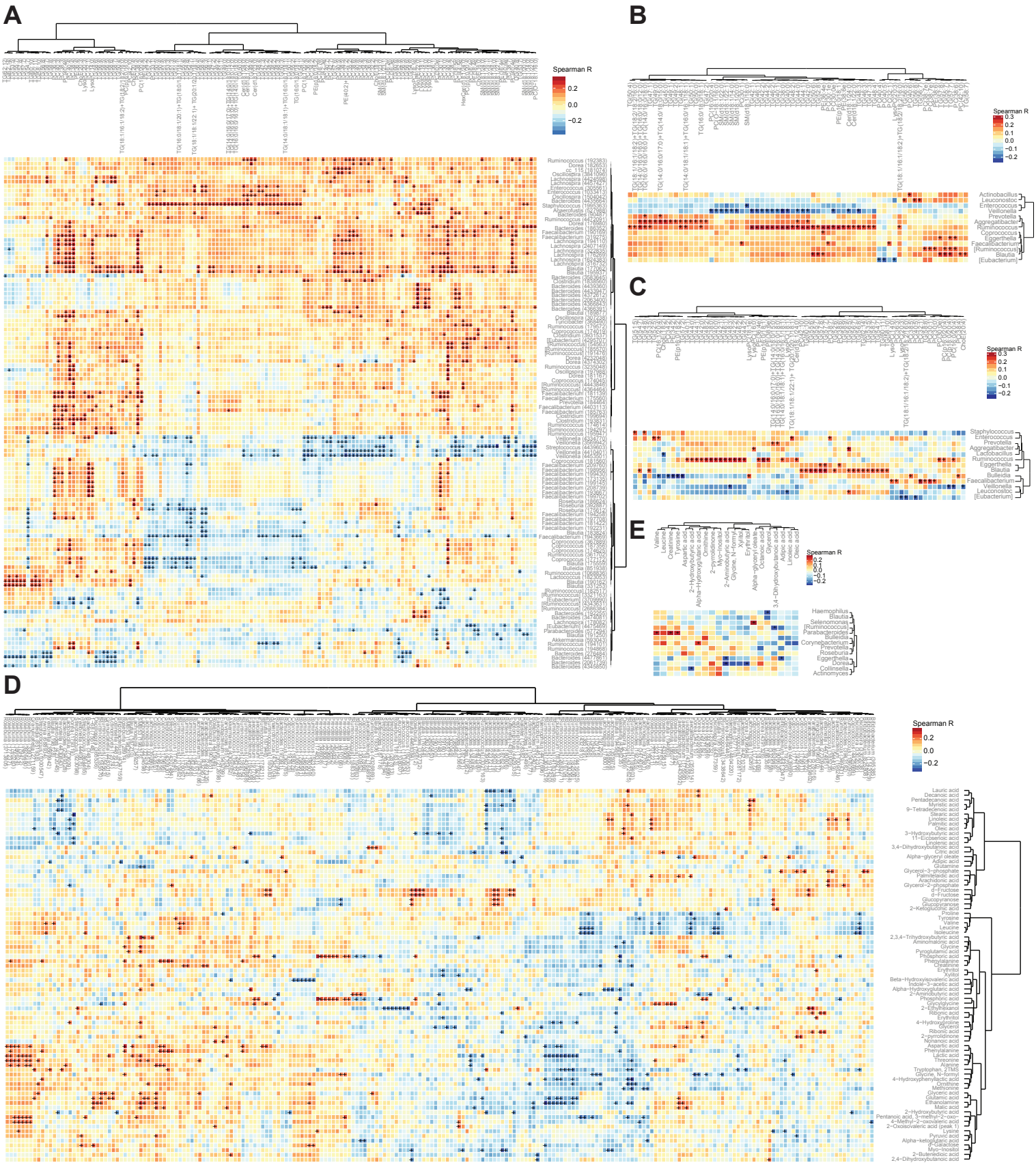

Figure S4, related to Figure 6. Spearman correlations between absolute abundances of metabolites and microbial relative abundances across all seven timepoints per individual. (**A-B**) Spearman correlations between serum lipids and gut microbiota 16S-based OTUs (A, showing the genus name and OTU ID for each) or genera (B). (**C**) Spearman correlations between serum lipids and gut microbiota 16S-based genera with time as a random effect after removing the random effect of time from the linear model. (**D-E**) Spearman correlations between serum metabolites and gut microbiota 16S-based OTUs (D) or genera (E). Significant correlations with $P < 0.01$ and $Q < 0.25$ indicated by a "+" and a "*", respectively.

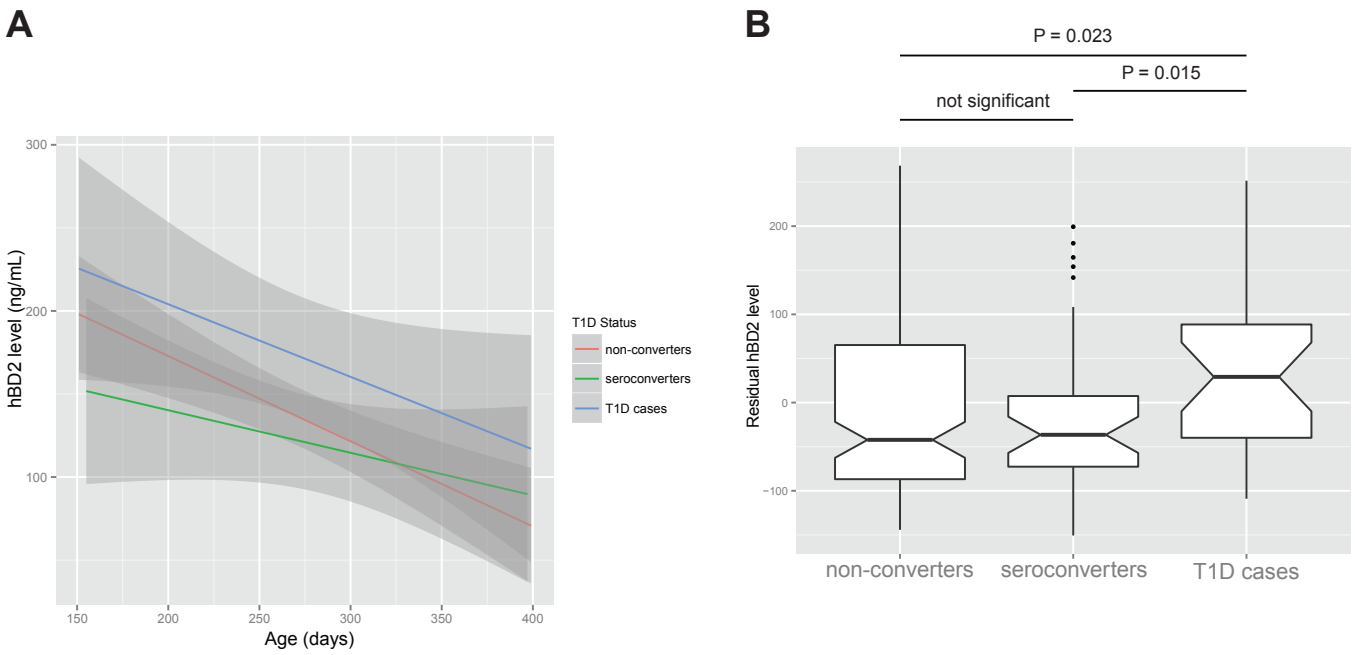

Figure S5, related to Figure 5, 6. Fecal human β-defensin 2 levels are elevated in T1D. (**A**) Fecal hBD2 levels for all children in the cohort are shown for all stool timepoints between 150 to 400 days, represented as a linear fit with confidence intervals (gray shading). (**B**) Boxplot of residual hBD2 levels after linear correction for age at sampling. Wilcoxon rank sum test with continuity correction for (i) non-converters vs. T1D cases: $P = 0.023$; (ii) seroconverters vs. T1D cases: $P = 0.015$; (iii) non-converters vs. serocoverters: $P = 0.98$.

Supplemental Tables

Table S1 – Detailed cohort information
Related to Figure 1 and Table 1

Table S2 – Complete 16S taxonomy data
Related to Figure 1 and Table 1

Table S3 – Strain level marker abundances
Related to Figure 4
Table S3A – Strain level marker recruitment
Table S3B – Species level marker coverage