Strain diversity of the human gut microbiome in early life

Tommi Vatanen[1], Damian R. Plichta[1], Juhi Somani[2], Timothy D. Arthur[1], Andrew Brantley Hall[1], Raivo Kolde[1], Moran Yassour[1], Kristiina Luopajärvi[3,4], Heli Siljander[3,4,5], Suvi M. Virtanen[6,7,8], Jorma Ilonen[9,10], Raivo Uibo[11], Vallo Tillmann[12], Sergei Mokurov[13], Natalya Dorshakova[14], Harri Lähdesmäki[2], Hera Vlamakis[1], Curtis Huttenhower[1,15], Mikael Knip, Ramnik J. Xavier[1,16,17,#]

[1]Broad Institute of MIT and Harvard, Cambridge MA, U.S.A.
[2]Department of Computer Science, Aalto University, Aalto Finland
[3]Children's Hospital, University of Helsinki and Helsinki University Hospital, Helsinki, Finland
[4]Research Programs Unit, Diabetes and Obesity, University of Helsinki, Helsinki, Finland
[5]Department of Pediatrics, Tampere University Hospital, 33521 Tampere, Finland
[6]Department of Health, National Institute for Health and Welfare, 00271 Helsinki, Finland
[7]School of Health Sciences, University of Tampere, 33014 Tampere, Finland
[8]Science Centre, Pirkanmaa Hospital District and Research Center for Child Health, University Hospital, 33521 Tampere, Finland
[9]Immunogenetics Laboratory, University of Turku, 20520 Turku, Finland
[10]Department of Clinical Microbiology, University of Eastern Finland, 70211 Kuopio, Finland
[11]Department of Immunology, Institute of Biomedicine and Translational Medicine, Centre of Excellence for Translational Medicine, University of Tartu, 50411 Tartu, Estonia
[12]Department of Pediatrics, University of Tartu and Tartu University Hospital, 51014 Tartu, Estonia
[13]Ministry of Health and Social Development, Karelian Republic of the Russian Federation, Lenin Street 6, 185035 Petrozavodsk, Russia
[14]Petrozavodsk State University, Department of Family Medicine, Lenin Street 33, 185910 Petrozavodsk, Russia
[15]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston MA, U.S.A.
[16]Gastrointestinal Unit, Center for the Study of Inflammatory Bowel Disease, and Center for Computational and Integrative Biology, Massachusetts General Hospital and Harvard Medical School, Boston MA
[17]Center for Microbiome Informatics and Therapeutics, MIT, Cambridge MA
[#] to whom correspondence should be addressed: xavier@molbio.mgh.harvard.edu

Abstract

The human gut microbiome matures towards the adult composition during first years of life. The microbial community assembly is affected by several intrinsic and extrinsic factors, including host genetics, living environment and human contacts. Here, we integrate the early gut microbiome data collected in DIABIMMUNE study in Finland, Estonia and Russian Karelia. We show that the gut microbiome is associated with linear growth, household location and elder siblings. Our SNP haplotype and metagenomic assembly based strain tracking reveal large and highly dynamic microbial pangenomes, especially in genus *Bacteroides*. We describe specific subspecies clades related to human milk oligosaccharide utilization and CRISPR system. Finnish and Estonian, but not Russian, microbiomes experience transient bloom of oral bacteria during infancy, whereas Russians commonly harbored a probiotic *Bifidobacterium bifidum* strain during the same time. This longitudinal study extends the current view of gut microbial community assembly on the strain level.

Introduction

Mounting evidence shows that the gut microbiome, particularly during its early development immediately after birth, plays an important role in human health[1,2]. Early childhood immune-mediated disorders including type 1 diabetes (T1D)[3,4], asthma[5,6], juvenile rheumatoid arthritis[7], allergic disease[8], and inflammatory bowel disease (IBD)[9] are linked to aberrations in the gut microbiota. Maturation of the immune system is orchestrated by early microbial exposures[10,11]; the complex relationship between the microbiome and the innate[12] and adaptive[13] immune systems during the first few year of life appears to be critical to later life health outcomes, but has not yet been explored at the population scale.

It has only recently become practical to investigate very detailed microbial exposures in early life at population scales. An increasing range of microbiome-linked health outcomes appear to be the consequence of individual strains of specific microbes[14-17]. These outcomes can result from structural variants in the gene products of individual strains[18], the presence or absence of gene cassettes from strains[19-21], or currently unexplained mechanisms. Until recently, most culture-independent methods appropriate for large-scale human populations (e.g., 16S rRNA gene amplicon sequencing) were limited in their ability to resolve such fine-grained differences. Now, both the efficiency of metagenomic sequencing and the availability of culture-independent strain-level analysis methods make more detailed investigation of the early life microbiome possible[22-24].

Complicating such studies, however, is the dynamic nature of the early gut microbiome and its numerous interactions with various intrinsic and extrinsic factors. While microbial exposures *in utero* are possible[1,25], the major colonization begins at birth when the neonate is either exposed to the vaginal microbiota during vaginal delivery[26] or to skin and environmental microbes after Caesarean section[27]. Subsequently, colonization is largely shaped by oligosaccharides and microbial constituents of human milk, a frequent cornerstone of the diet in infancy[28-30]. The

assembly of microbial communities is further influenced by the introduction of solid foods, use of antimicrobials, host genetics, geography, and numerous other environmental factors[31].

In T1D, several human cohort studies have reported alterations in the gut microbiota[32-38] and increased intestinal permeability[39] prior to diagnosis, but mechanisms connecting gut health to destruction of pancreatic beta cells remain unknown. One such cohort, DIABIMMUNE, aimed at identifying microbial factors implicated in T1D and preceding islet autoimmunity. DIABIMMUNE includes nearly 700 children, selected by their human leukocyte antigen (HLA) haplotypes conferring increased risk to autoimmune disorders, in three neighboring countries: Finland, Estonia, and Russian Karelia. These children were observed for three years from birth by monthly stool sampling, frequent questionnaires about common life events and circumstances, and periodic blood sampling to track different immune parameters.

The DIABIMMUNE longitudinal cohort study enables in-depth characterization of the developing gut microbiome and various immune markers in the context of T1D. Within DIABIMMUNE, a case-control study of children with T1D or beta cell autoimmunity (n = 11 cases) found decreased microbial diversity and an increase in inflammation-favoring organisms preceding the diagnosis of clinical T1D[40]. Another investigation showed a decrease in microbial diversity and an increase in antibiotic resistance genes in connection with recurrent antibiotic treatments[41]. Two additional DIABIMMUNE studies underscored the importance of microbial lipopolysaccharide (LPS) exposures *in utero*[42] and during the first years of life[43]. The latter study also correlated differences in LPS structure with immunogenicity, providing a mechanistic link between LPS subtypes and T1D[43]. Epidemiological investigations established recurrent early life infections as a risk factor for beta cell autoimmunity, T1D, and celiac disease[44,45], and found an association between rural living environment (forest and agricultural land) and decreased atopic sensitization[46]. A study analyzing circulating cytokines in peripheral blood samples established a connection between upregulation of the IL-17 pathway and advanced beta cell autoimmunity[47].

Here, we set out to further characterize the early gut microbiome using an integrated and extended dataset from DIABIMMUNE consisting of 16S rRNA gene sequencing of 3,204 samples and metagenomic sequencing of 1,154 samples, together spanning 289 subjects at an average of 11.4 (range 1-36) time points per subject. Using the 16S data, we demonstrate that multiple intrinsic and extrinsic factors, including household location (urban vs. rural), early growth, elder siblings, and multiple maternal factors are associated with features of the early gut microbiome. We conduct in-depth strain identification and characterization using both SNP haplotyping and metagenomic assembly. Applying these two methods to this longitudinal data set, we describe strain acquisition, diversity and interactions, and pangenome dynamics of common early gut species with specific examples from genera *Bifidobacterium*, *Bacteroides*, and others. Taken together, the integrated DIABIMMUNE microbiome data provide detailed, strain-level characterization of the developing gut microbiome.

Results

The DIABIMMUNE study followed children from Finland, Estonia, and Russia for three years starting from birth by collecting monthly stool samples, periodic serum samples, and frequent questionnaires on early life events. Here, we set out to integrate all published microbiome data that have been generated in multiple DIABIMMUNE studies using both 16S rRNA amplicon and metagenomic sequencing techniques[40,41,43]. After quality control, the data consisted of 3,204 16S amplicon and 1,154 metagenomic sequencing profiles from 289 and 269 study subjects, respectively (**Table 1, Fig S1**).

	Espoo, Finland	Tartu, Estonia	Petrozavodsk, Russia
Study subjects	139	78	72
16S sequencing samples (median per subject)	2080 (9)	501 (6)	623 (7)
Metagenomic sequencing samples (median per subject)	616 (4)	221 (3)	317 (3)
Males	77	39	40
Females	62	39	32
Caesarean sections	9	6	12
Maternal age at birth, mean (sd)	31.1 (5.0)	29.3 (5.1)	27.8 (4.7)
Born in rural household	10 (7.8 %)	18 (23.1 %)	0
Median number of elder siblings (range)	1 (0-4)	1 (0-4)	0 (0-2)
T1D AAB seropositive subjects	11	4	1
Subjects with T1D diagnosis	5	1	1

Table 1. DIABIMMUNE microbiome cohort statistics. Distribution of study subjects, stool samples with sequencing data and several other external variables across the study sites. Table shows number of study subjects (N) per category unless otherwise specified. T1D autoantibody and diagnosis information as of Nov 2016.

Early life events are reflected in the gut microbiome

External factors such as household location (city vs. countryside), daycare attendance, and elder siblings can affect microbial exposure, but less is known about the actual impact of these factors on gut microbial communities[48,49]. Additionally, maternal variables such as antibiotic courses during pregnancy and maternal age may directly affect the microbiome of the infant. To extend the understanding of early microbial development in connection with external variables, we first analyzed the more ample 16S data (n = 3,204 samples) using both omnibus and individual association tests. By cross-sectional Permutational analysis of variance (PERMANOVA), we found that in addition to well-known features affecting the early microbial composition (birth mode, geographic location, and breastfeeding status), maternal antibiotic course(s) during pregnancy (permutation test, q-value = 0.029), and maternal age at birth (q-value = 0.16) were associated with microbial composition shifts in the earliest stool samples collected at 2 months of age (**Table S1**). While the effect of maternal antibiotics was seen only in the earliest cross-section, maternal age continued to show borderline significance at month 6 (q-value = 0.14, **Table S1**).

We next associated the gut microbial diversities (Chao1 richness, Shannon's diversity index) with the above-mentioned external factors and observed associations with age of sample collection, breastfeeding, geography and antibiotics consistent with previous studies (**Table S2**)[41,50,51]. Additionally, height at age three (linear mixed effects model, q-value = 0.097) and growth rate (average increase in height per year) during the first three years (q-value = 0.10) were associated with microbial diversity; taller and faster growing children had higher diversity trajectories throughout the three year follow-up (**Fig. 1A**), suggesting a link between the gut microbiome and growth in early childhood. Children living outside cities harbored more rich microbiomes compared to children in urban households throughout the first three years of life (q-value = 0.025, **Fig 1B**), confirming that microbial exposures from rural environments are directly reflected in the gut. On a taxonomic level, the weight at the age of three (q-value = 0.0047) and weight gain during first three years (q-value = 0.00080) were positively correlated with the relative abundance of genus *Dialister* (**Fig. S1C**). Finally, Finnish and Estonian subjects with elder siblings tended to have more *Bifidobacterium* spp. in their early samples (q-value = 0.15, **Fig S1B**, **Table S3**), possibly due to lateral transfer from elder siblings. The findings of our association analyses, summarized in **Table S1, S2** and **S3**, contribute to the understanding of early microbial colonization and community assembly in the human gut.

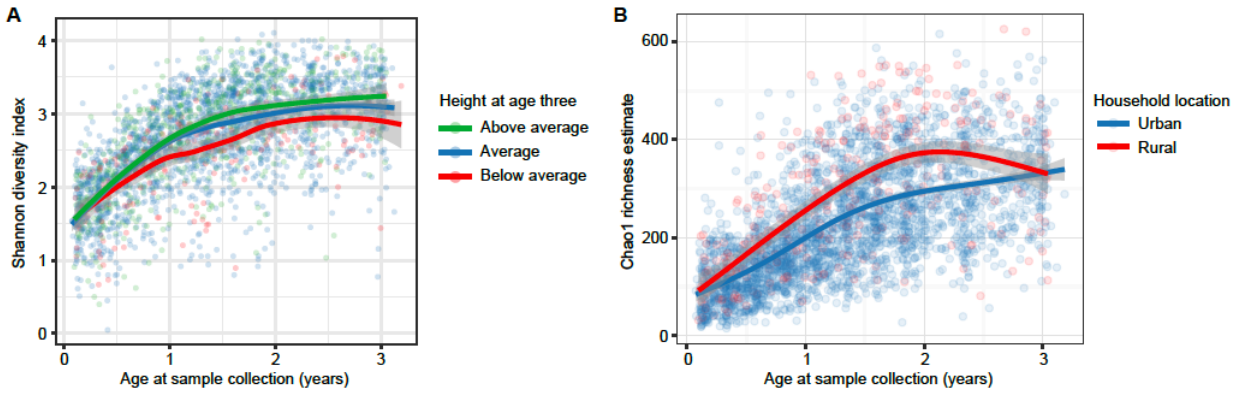

Figure 1. Associations between early gut microbiota and intrinsic and extrinsic factors. A Children's height at the age of three is correlated with gut microbial diversity (q-value = 0.14). Weight categories were defined as follows; above average: weight z-score > 1, average: -1 < weight z-score ≤ 1, below average: weight z-score ≤ -1,**B** Children born in rural households harbor more diverse gut microbiota (q-value = 0.068).

Strain diversity and ecology in the early gut

To expand the analysis beyond the typical for 16S data genus level, we leveraged shotgun metagenomic data to perform in-depth strain-level analysis in terms of both single nucleotide polymorphisms (SNPs) or gene content. Strain analysis has the potential to delineate novel microbial sub-populations[52,53] and to identify potential functional adaptations in the gut microbiome[54,55]. Particularly, *de novo* strain identification is important for microbial species with a limited number of isolated strains, and the gut microbiome has many such understudied species despite large cultivation efforts[56,57]. We first profiled our metagenomic sequencing data (n = 1,154 samples) by strain haplotyping on species-specific conserved and unique marker genes[23], which identified the dominant strain for the most abundant species in every sample. We then compared the resulting SNP haplotypes by sequence similarity and stratified them in intra- and inter-subject comparisons (**Fig. 2A, Table S4**). Longitudinal, intra-subject comparisons showed more similar strains compared to inter-subject comparisons, consistent with previous observations[52,58,59]. Overall, we found a wide range of strain diversities among investigated bacterial species (**Fig. 2A, Table S4**). For example, *Haemophilus parainfluenzae* and *Faecalibacterium prausnitzii* were among the most diverse species, with strains that had less than 95 % sequence similarity in SNP haplotype comparison. On the contrary, all investigated members of genus *Bacteroides* had very low sequence variability, accentuated by virtually identical SNP haplotypes in intra-subject comparisons (mean sequence similarity 99.96 %) and accompanied by, on average, greater than 99.6 % sequence similarity in inter-subject comparisons. All other species analyzed had an average inter-subject similarity of 98.9 %.

The observed high level of sequence identity in *Bacteroides* spp. contradicted existing evidence of their genome diversity in terms of gene content[60]. This led us to speculate that the SNP-based evaluation we performed did not reflect all facets of genetic diversity of the gut microbes, whose evolution is shaped by lateral gene transfer (LGT)[61], large effective population size (N_e)[62] and niche adaptation[63,64]. To investigate one *Bacteroides* species in detail, we isolated and sequenced eight *Bacteroides dorei* strains - three from two DIABIMMUNE stool samples

(including two different isolates from a single stool sample) and five from adult stool samples - using PacBio long read sequencing. This data enabled trivial assembly of high quality genomes, and when merged with seven existing NCBI isolate genomes, it expanded the pangenome (the collection of genes or gene families found in the genomes of a given species) of *B. dorei* by 7,828 genes to almost 18,000 unique gene families (**Table S5**). Interestingly, each newly sequenced isolate genome harbored between 276 and 1,168 (median 750) unique accessory genes, which on average represented 13 % of the genes in each *B. dorei* strain (**Table S5**). For all 15 *B. dorei* isolates, this variability translated to 70 % inter-strain similarity on the gene content level, on average, which is considerably lower than the observed SNP based similarity (**Fig. S2C)**. Each of the newly-sequenced strains encoded between 17 and 63 accessory gene islands (regions consisting of contiguous accessory genes) that were significantly longer compared to randomly permuted data (>15 genes, $P < 0.01$) (**Fig. S2D**). Five of these were encoded on contigs that could be circularized, suggesting an episomal entity (likely a plasmid) (**Table S6**). Anecdotally, the *B. dorei* strains isolated from the DIABIMMUNE samples were more similar to each other, having 91 % similarity between isolates from the same individual and 83-89 % similarity between isolates from different infants from different countries (**Fig. S2C**). In comparison, *B. dorei* isolates from adults in the PRISM cohort were on average 68 % genetically similar. This indicates that later in life the gut microbiome is inhabited by more genetically diverse *B. dorei* strains, likely reflecting more heterogeneous dietary regimes and lifestyle[65]. Whether that happens through evolution and/or adaptation of early colonizers or strain replacement remains unclear.

To investigate accessory genes across all taxa, we turned to *de novo* metagenome assembly of the DIABIMMUNE metagenomic samples. This expanded the gene pool (number of observed gene families) to 6,328,944 non-redundant genes, compared to 1,932,010 gene families found using pangenomes constructed from the NCBI isolate genomes[66]. We binned the assembled metagenomes using a co-abundance technique[67] into metagenomic species and constructed pangenomes for 22 species from these metagenomic assemblies (**Fig. 2B-C**). The *de novo* assembled pangenome of *B. dorei* consisted of roughly 28,000 genes and included 93 % of the genes identified in the isolate sequences from DIABIMMUNE samples and 82 % of genes identified in the remaining PRISM isolate sequences. The lower recall rate (sensitivity) of the latter reflects that PRISM metagenomes (or any other adult samples) were not included in the *de novo* assembly and suggests that the constructed pangenome of *B. dorei* was not fully saturated and will be extended by using additional metagenomes or isolates from different populations; we expect the same trend to be true for other species as well. Among all analyzed species, *Bacteroides* spp. and *E. coli* harbored the largest assembled pangenomes, each with more than 25,000 genes (**Fig. 2C**). This is concordant with high genome plasticity observed in *Bacteroides*[60], whereas *E. coli* is an omni-present species with more than 60,000 gene families in its NCBI isolate pangenome[68]. Consistent with the SNP haplotype analysis, the strains recovered from the same individual (intrasubject comparisons) were more similar to each other than the strains found in different individuals (intersubject comparisons, **Fig. 2B, Table S4**). In contrast, the magnitude of variability was much higher in terms of gene content compared to SNP haplotypes (**Fig. 2A, B**). These two measures were highly correlated in most species but showed low or no correlation in a minor subset of species, including *F. prausnitzii* and *B. dorei*

(**Fig. 2D**). In *B. dorei*, the results from metagenomic assemblies and isolate genomes showed a similar trend suggesting that the lack of correlation between the SNP haplotypes and assembled genomes was not an artefact of the metagenomic assembly (**Fig. 2E**, orange points). Rather, it indicates more rapid (or slow but high in volume) diversification in the accessory gene content compared to the pace of acquisition of random point mutations in the core genome. In contrast, *E. coli* metagenome assemblies displayed a high correlation between gene content and SNP haplotype similarities (**Fig. 2F**) in agreement with established notions.

We again used the longitudinal nature of our study to compare the difference between strains from the same or from different individuals in their gene content. On average, the metagenomic assemblies of *B. dorei* had 86 % gene content similarity in intrasubject comparisons and significantly lower (76 %) similarity in intersubject comparisons. Notably, these values fall into the same range as we observed in comparisons between *B. dorei* isolate genomes (**Fig. S2C**). *H. parainfluenzae* was an outlier with a very similar measure of within- and between-subject gene similarity at 66 % and 64 %, respectively. This may reflect transient gut colonization events and frequent replacement with new strains descending from the oral cavity where *H. parainfluenzae* is autochthonous, as we discuss in detail below. When comparing several species of *Bifidobacterium*, we observed a greater variability in gene content and SNPs in the intrasubject comparisons for *B. longum*, relative to *B. bifidum* or *B. breve* (**Fig. 2A, B**), leading us to explore the functional consequences of this strain-level variation in more detail in the following section.

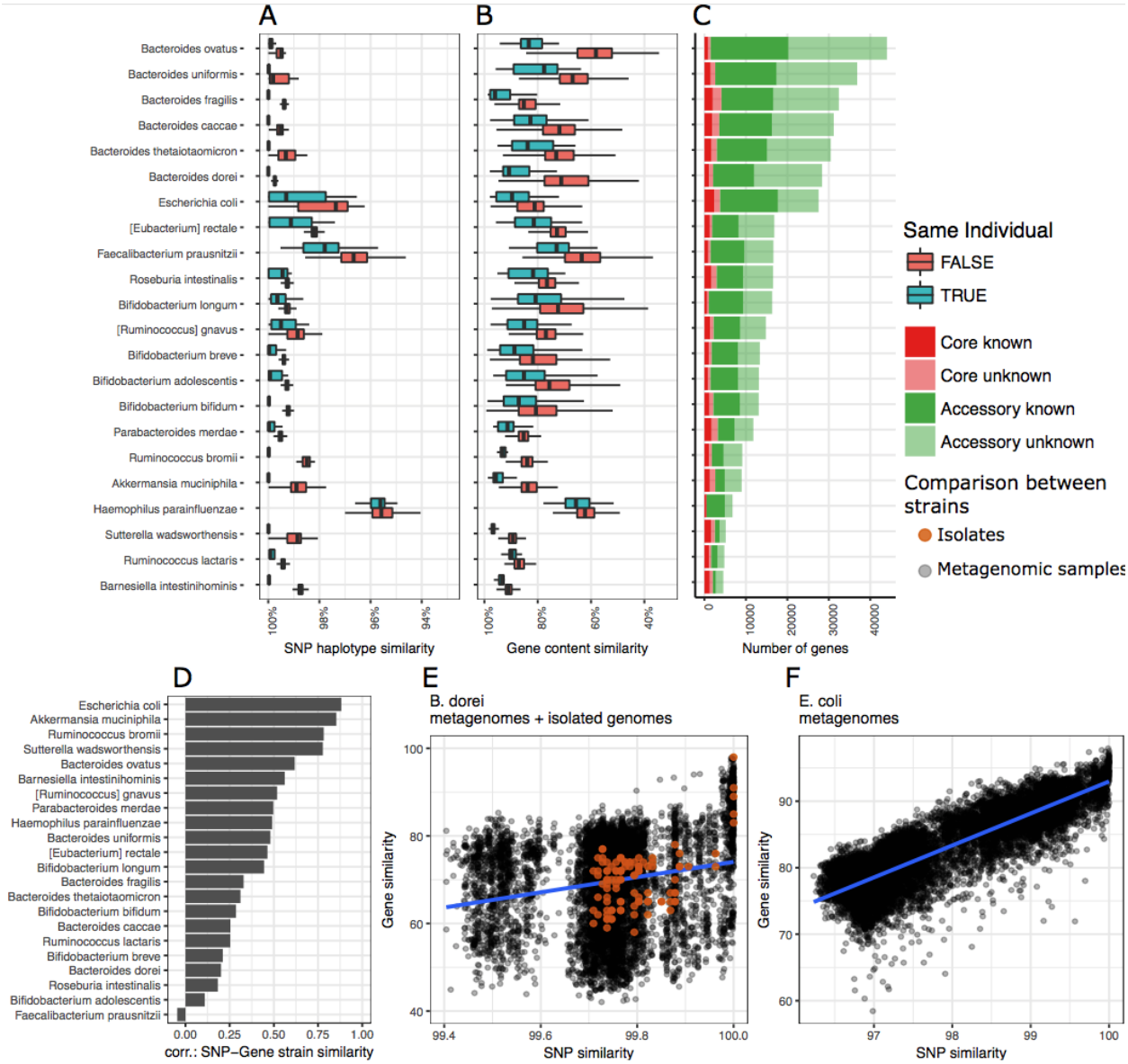

Figure 2. Strain diversity across species in early gut metagenomes. A SNP haplotype similarities per species based on all pairwise comparisons (dominant strain per species per sample) and stratified to intra-subject and inter-subject comparisons. Species containing >10 comparisons in both strata are shown, and the order is given by increasing median similarity. The box shows the interquartile range (IQR), the vertical line shows the median and the whiskers show the range of the data (up to 1.5 times IQR). **B** Gene content similarities per species, evaluated on pangenomes generated by metagenomic assembly. Boxplots as in panel **A**. **C** The size of core and accessory genomes per species stratified by the functional annotation of genes using eggNOG #### and ordered as in panel **A**. **D.** Correlation between SNP based and gene content based similarity between strains. **E.** *B. dorei* strains' similarity at SNP and gene content levels are uncorrelated (r=0.1). Comparisons between isolated genomes are shown in orange for reference. **F.** *E. coli* strains' similarity at SNP and gene content (r=0.86).

Strains in *Bifidobacterium* spp. reflect breastfeeding patterns and geography

Bifidobacteria are widely-characterized beneficial gut commensals, commonly dominating the gut during breastfeeding and later dissipating throughout life. They possess immunomodulatory functions, produce beneficial metabolites (e.g., vitamins, extracellular polysaccharides, and short-chain fatty acids), and metabolize a wide range of diet-derived, nondigestible carbohydrates (e.g., oligosaccharides, polyols, and dietary fibers)[69]. Specifically, the subspecies found in the infant gut typically harbor a wide variety of genes that enable the use of human milk oligosaccharides (HMOs) as the sole energy source[70]. *B. longum* subsp. *infantis* (*B. infantis*)[71] and some strains of *B. longum* subsp. *longum*[72] are capable of membrane transport and intracellular degradation of intact HMOs, whereas other subspecies in the *B. longum* clade rely partially on extracellular enzymes for HMO utilization[73]. To identify different *B. longum* subspecies in the metagenomic data, we surveyed the metagenomes for the genes of a well-characterized cluster responsible for HMO transport and degradation in *B. infantis*[71]. The presence of these genes was strikingly consistent with the SNP haplotype-based phylogeny of *B. longum* strains (**Fig. 3A, B**). Two *B. infantis* reference sequences (ATCC 15697) clustered with 70 strains harboring these genes (highlighted in **Fig. 3A**) observed in the metagenomes. We found evidence for the presence of this gene cluster in 14 additional samples. In these cases, it is possible that these communities harbored multiple *B. longum* strains, of which *B. infantis* is non-dominant, and that the SNP haplotype profile was not based on *B. infantis*. Comparing the communities with *B. infantis* (defined by presence of the HMO gene cluster) to communities with other *B. longum* strains revealed evidence of a competitive advantage that allows *B. infantis* to reach higher relative abundances on average than other *B. longum* strains (**Fig. 3B,** linear mixed effects model $p = 0.00049$), albeit with modest effect sizes.

Commercial strains of different *Bifidobacterium* spp. are commonly used in probiotic supplements and foods. One such species, *B. bifidum*, showed contrasting relative abundances between the countries: unlike Finnish and Estonian samples, early Russian samples commonly contained more than 10 % of *B. bifidum* (**Fig. 3C**). Investigating the SNP haplotypes of *B. bifidum* revealed that 79 samples from 34 Russian, 3 Estonian, and 2 Finnish subjects harbored the same *B. bifidum* strain with greater than 99.9 % sequence similarity (**Fig. 3D**). This SNP haplotype was identical to the NCBI isolate genome *B. bifidum* 791, which was isolated from a healthy human gut in 1993 in Nizhny Novgorod, Russia and has been patented for medical use in Russia. *B. bifidum* relative abundance was greater than 10 % in 57/79 (75 %) samples containing this strain. Together with repeated detection of this strain from 17 subjects (with a maximum of seven observations from two Russian subjects), these observations suggest that *B. bifidum* 791 attained stable engraftment in these gut communities.

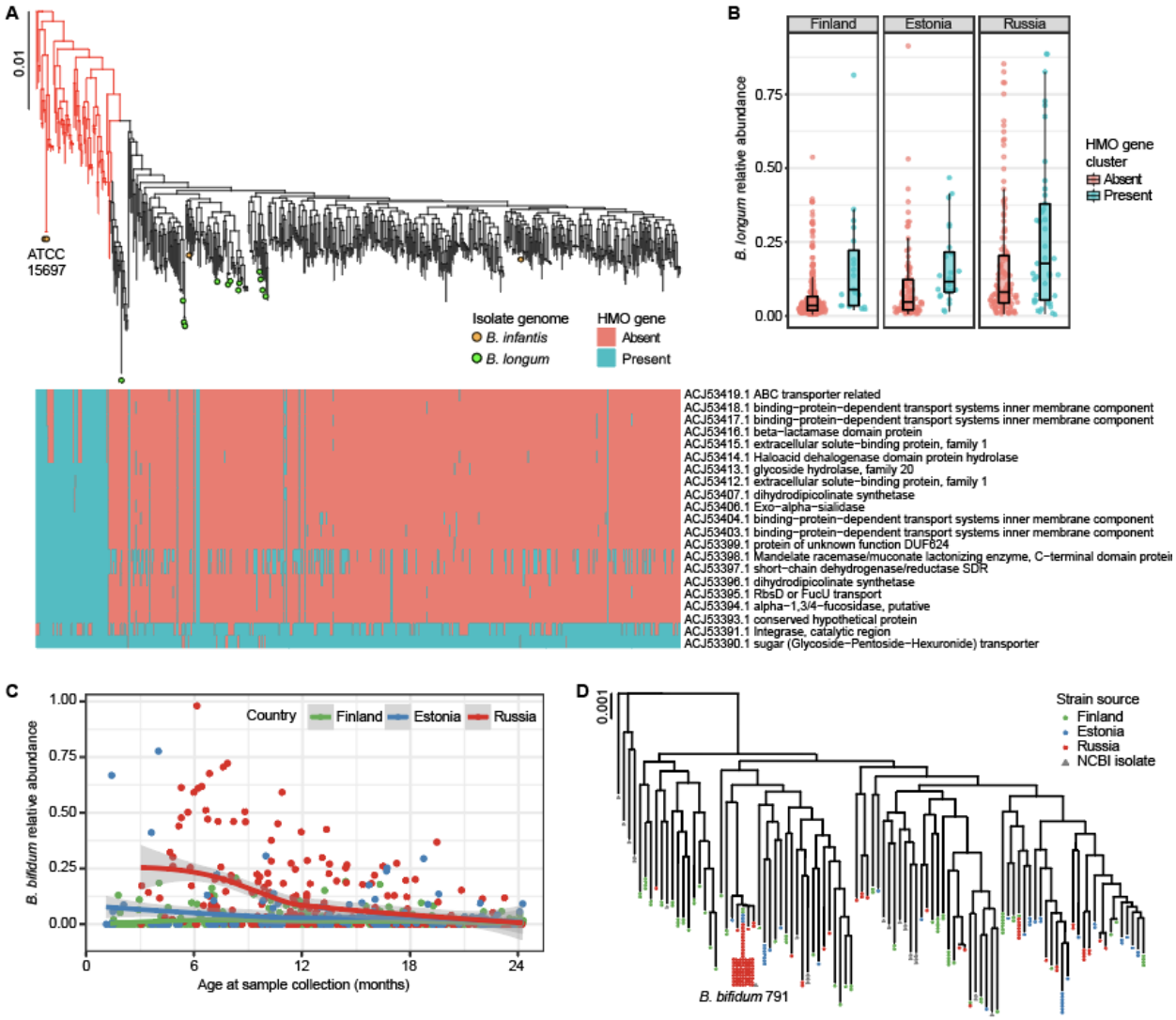

Figure 3. Bifidobacterium strains in DIABIMMUNE children. A Phylogenetic tree of *B. longum* strains in DIABIMMUNE stool samples together with 18 NCBI *B. longum* isolate genomes based on SNP haplotypes. The heatmap illustrates strain-specific carriage of 21 genes in *B. infantis* HMO gene cluster, responsible for intracellular HMO degradation, evaluated using the metagenomic data. Highlighted strains include two reference sequences of *B. infantis* (ATCC 15697). **B** Boxplot of *B. longum* relative abundance stratified by country (facet) and *B. longum* strain; *B. infantis* (highlighted in panel **A**) vs. other *B. longum* strains. **C** Relative abundance of *B. bifidum* longitudinally stratified by the countries. Russians have more B. bifidum especially during the first year of life. **D** Phylogenetic tree of *B. bifidum* strains in the DIABIMMUNE stool samples based on SNP haplotypes. Strains with >99.5% sequence similarity have been collapsed into a single tip. A known strain, *B. bifidum 791*, was found in 79 stool samples.

Oral strains appear transiently in infant gut

To more broadly contextualize the gut bacteria in DIABIMMUNE and to compare the developing gut microbiome with established adult microbiomes, we compared the strains in this study with the strains of healthy adults in the Human Microbiome Project (HMP) study[52]. In addition to the gut, HMP obtained metagenomic data from three other major body areas: skin, oral cavity, and vagina. We first stratified the species observed in the DIABIMMUNE gut samples into four categories by their typical habitat in HMP. Each bacterial species was assigned to one of four

habitats (adult gut, skin, oral cavity, or vagina) by the highest mean relative abundance in HMP data (**Fig. S3A, Table S7**). By applying these strata to DIABIMMUNE samples, we saw an increasing abundance of adult gut bacteria with age at sample collection that reflected maturation of microbial composition (**Fig. S3B**). There was a reciprocal longitudinal dissipation of vaginal and skin bacteria (**Fig. S3C,D**), which were commonly seen in higher abundances during the first months of life. Notably, oral bacteria spiked during the first year of life in Finnish and Estonian infants (**Fig. 4A**). Bacteria in this strata included common opportunistic pathogens (pathobionts) such as members of genera *Veillonella*, *Haemophilus,* and *Streptococcus* (**Table S7**), many of which have also been isolated from the upper gastrointestinal tracts of elderly adults[74]. In Russian infants, the migration of these oral bacteria may be prevented by higher levels of *Bifidobacterium* spp. in the gut, which provide colonization resistance against such opportunistic bacteria[75]. This may also partly explain the differences in infant immune development between the countries in this study, as colonization of oral bacteria has been shown to drive Th1 cell induction and inflammation[54].

Some bacterial species, including the oral taxa *Veillonella parvula* and *Haemophilus parainfluenzae* that had the highest mean relative abundance in DIABIMMUNE subjects, consist of distinct, body site-specific subspecies clades[52]. To examine how these clades were related to the strains appearing in the infant guts, we integrated the metagenomic strain SNP haplotypes with the HMP data. *V. parvula* strains in infant guts were similar to oral strains found on buccal mucosa and dental plaque but distinct from a more diverse clade typical of tongue microbiome (**Fig. 4B, S3E**). Conversely, the variability of the infant gut strains of *H. parainfluenzae* spanned HMP tongue and buccal mucosa strains but tended to be distinct from adult dental plaque strains (**Fig. 4C, S3F**). In infants, genera *Veillonella* and *Haemophilus* have been associated with formula feeding and different human milk oligosaccharide structures[76,77]. These observations demonstrate strain level differences in oral bacteria colonizing the infant gut in relation to the adult oral microbiome.

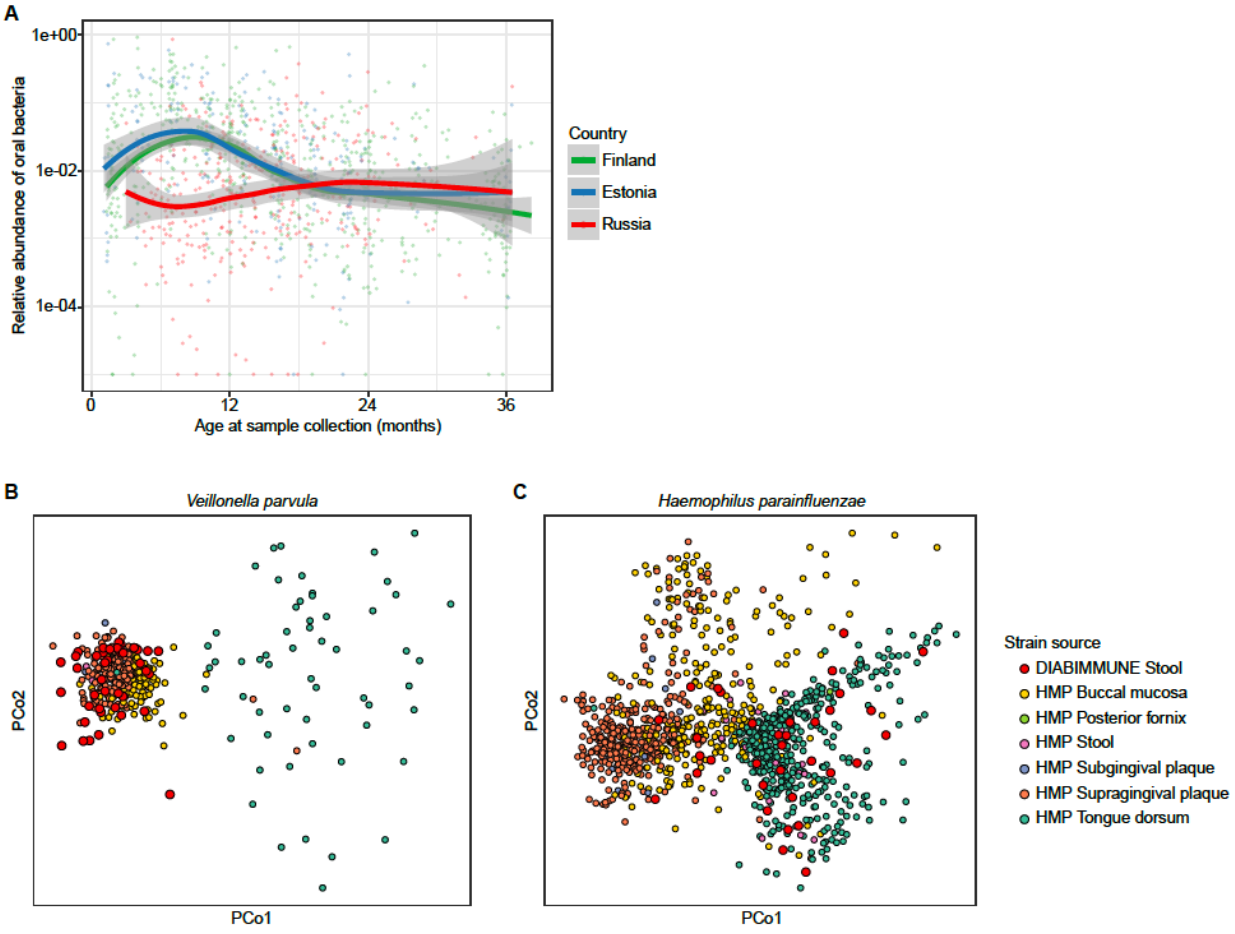

Figure 4. Finnish and Estonian infant gut harbor oral bacteria. A Relative abundance of oral bacteria (defined by highest mean relative abundance in HMP oral samples, compared to other body areas) in DIABIMMUNE metagenomes. Oral bacteria bloom in Finnish and Estonian infants during the first year of life. **B-C** Ordinations of (**B)** *Veillonella parvula* and (**C)** *Haemophilus parainfluenzae* strains appearing in DIABIMMUNE infants' guts together with the strains in HMP adults exhibiting body site specific subspecies clades.

Ruminococcus gnavus clades differ by CRISPR genes

Differences in CRISPR system carriage by human-associated microbes is important both due to their potential roles in speciation, and as a route to genetic intervention during controlled experiments[78-80]. Roughly half of known bacterial genomes harbor CRISPR system[81] which led us to wonder whether there are any strain level differences in CRISPR system carriage within species. To this end, we first screened the metagenomic data for gene families involved in CRISPR system. We specifically focused on 50 most prevalent and abundant species in this data and calculated the prevalence of each CRISPR gene per species in samples where the relative abundance of the species in question was greater than 5 %. We further focused on genes with prevalence between 25 % and 75 % (indicating that only a limited subset of strains within species carried these genes) and found 123 gene families across 26 bacterial species (**Table S8**). These genes exemplify tentative speciation or niche adaptation in many gut

commensals, such as Bacteroides species *B. vulgatus* and *B. fragilis* and *Bifidobacterium breve*.

To further investigate such adaptation in detail, we focused on *R. gnavus* which tentatively harbored six contrasting CRISPR genes and showed a bimodal SNP haplotype similarity distribution, implicative of two distinct clades or subspecies (**Fig. S2A**). Phylogenetic analysis of the SNP haplotypes revealed two distinct subspecies clades as previously described in adult IBD patients (**Fig. 5A**)[16]. To survey any functional differences between these two *R. gnavus* clades, we compared the prevalence of genes in an extended *R. gnavus* pangenome[16] between the clades (**Table S9**). Among the genes with differential prevalences between the clades we found seven genes involved in the CRISPR system and additional genes related to phage activity and drug resistance (**Fig. 5B**). These differences provide evidence for phage associated niche adaptation in formation of clade 1 (blue) harboring the CRISPR genes. Such adaptation might occur under conditions abundant with *R. gnavus*- targeting phages through gradual habitat filtering (strains with any resistance against phages have colonization advantage), while the other clade (clade 2, red) might lose (or not obtain) CRISPR genes in the absence of similar pressure.

Within the DIABIMMUNE cohort, 39 non-IBD subjects harbored a strain belonging to clade 2, which has been previously found only in adults with IBD[16]. Anecdotally, clade 1 that was present in the majority (92/152) of the DIABIMMUNE samples was also found in 15 samples from four children with clinical T1D diagnosis, whereas clade 2 was present in only a single sample from one T1D-positive child who harbored clade 1 in a separate sample.

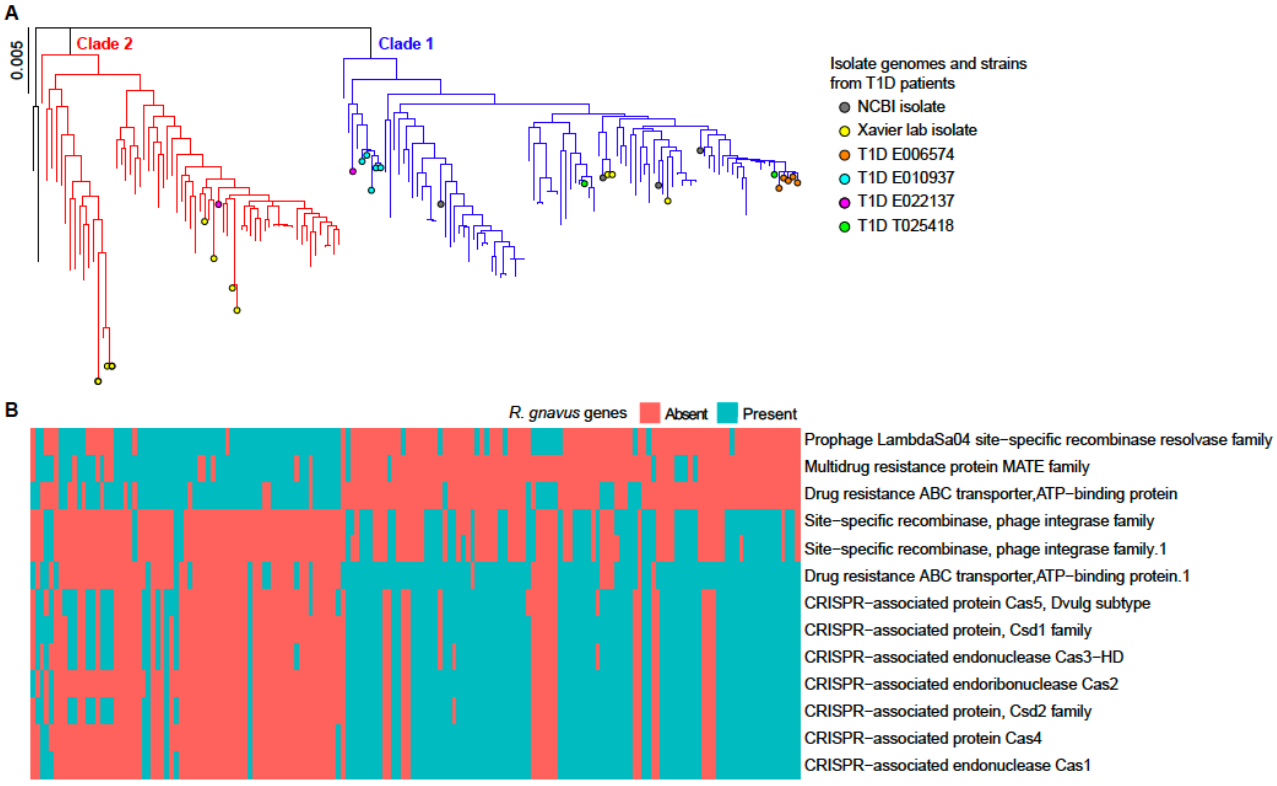

Figure 5. *Ruminococcus gnavus* clades in DIABIMMUNE gut samples. A Phylogenetic tree of *R. gnavus* strains in DIABIMMUNE samples, with existing isolate genomes and samples from children diagnosed with T1D during follow-up annotated with circles: Xavier lab isolates were described in[16]. **B** Presence of selected genes with differential prevalences between the clades. See **Table S7** for a complete list of genes and their prevalences per clade.

Contributional diversity of microbial functions

Finally, we turned from the species-centric to function-centric view of the microbiome. We binned species into functional 'guilds' based on the functional pathways they share to assess their contributional diversity, i.e. to assess how diverse set of species encode and have a potential to perform a given function per sample[82]. We assessed the contributional diversities for 365 Gene Ontology (GO)[83] biological process terms present in more than 100 metagenomes. Unsurprisingly, most GO terms displayed increasing within-sample functional diversity (Gini-Simpson index) with increasing age that coincides with microbiome maturation and increasing diversity (**Fig. 6A**, **Table S9**). Many widely distributed pathways such as sporulation (GO:0030435), glycolysis (GO:0006096), and riboflavin (vitamin B2) biosynthesis (GO:0009231) followed this pattern. In contrast, a few specific pathways did not display this increasing trend; aerobic electron transport chain ($r = -0.16$, $q = 0.001$, GO:0019646), viral release from host cell ($r = -0.05$, $q = 1.0$, GO:0019076), and siderophore biosynthetic process ($r = -0.07$, $q = 1.0$, GO:0015891) showed decreasing or stable trends in time (**Fig 6A**, **Table S10**).

Between-sample contributional diversities (beta-diversity, Bray-Curtis dissimilarity) reflect the stability of functional contributions per pathway and can be assessed longitudinally within and between subjects. We observed a decreasing trend in contributional beta-diversities with increasing age (**Fig. 6B**, Pearson $r = -0.28$, $p < 2.2e-16$), reflecting an overall maturation and stabilization of the microbiome. Microbial contributions to pathways were more stable within individuals (Student's *t*-test $p < 1e-20$ in all time windows), as reflected by lower beta-diversities, and the gap between within- and between-subject comparisons tended to widen with time similar to the average beta-diversities of taxonomic profiles (**Fig. 6B**). This view provides another perspective of the early stabilization of gut microbial communities: as pathways in some cases reflect ecological niches (e.g., aerobic electron transport), the above trend may mirror convergence to specific ecological attractor states, which in turn results in a stable state after community adaptation and competition over the niche has resolved (**Fig. 6B**).

Pathways related to bacterial acquisition of iron by siderophores highlighted in **Fig. 6A** provide an example of how to interpret contributional diversities. Bacteria secrete iron-binding siderophores to harvest iron, but extracellular siderophores are exploited by other bacteria. According to the black queen hypothesis, the ability to produce such costly but necessary molecules is under negative selection until the production is minimal but sufficient to support the microbial community[84]. Indeed, according to our data, siderophore biosynthesis is contributed by a single dominant species per community (i.e., low contributional alpha-diversity, **Fig. 6A,C**) whereas siderophore transport-related genes are more widely distributed across the community members (**Fig. 6A,D**).

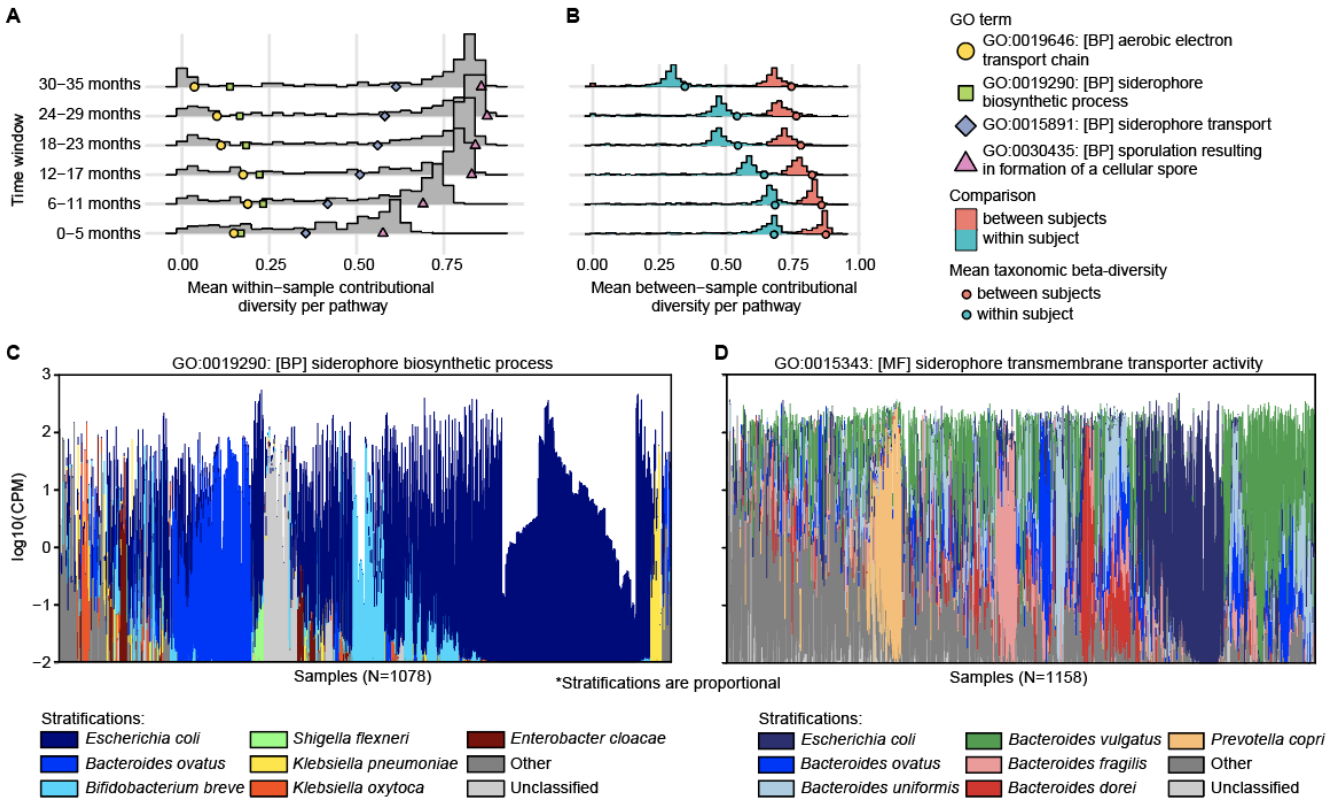

Figure 6. Contributional diversity of microbial pathways. A-B We applied alpha- (**A**) and beta-diversity (**B**) to the distribution of species contributing to functional categories, GO biological process terms, measuring their contributional diversities. The histograms show the mean alpha (**A**) and beta-diverities (**B**) per GO term stratified in time windows on y-axis. Colored points show (**A**) examples of pathways with different trends, and (**B**) mean intra- and intersubject beta-diversities of taxonomic profiles. **C-D** Species contributing to (**C**) siderophore biosynthetic process and (**D**) siderophore transport. Colors displaying the contributions of individual species are linearly scaled within the log-scaled total bar height depicting the total abundance of the pathway.

Discussion

We reported a longitudinal, strain-level investigation of the developing gut microbiome utilizing the DIABIMMUNE cohort and its rich metadata of various life events. We found associations between microbial features and early linear growth, household location, maternal antibiotic courses and elder siblings. Our SNP haplotype and metagenomic assembly-based analyses revealed that many common gut commensals, such as *Bacteroides* spp., have large and dynamic pangenomes with tens of thousands of genes. We showed that *B. infantis*, which is highly specialized in human milk utilization, has a competitive advantage over other *B. longum* strains in early gut microbiota in this uncontrolled, prospective setting. A commercial strain of another common Bifidobacterium species, *B. bifidum*, was commonly dominating early Russian microbiomes. These analyses contribute to taxonomic and functional understanding of early gut communities.

Our analyses revealed associations between microbial features and early linear growth. Height (but not weight nor body-mass-index, BMI) at age three and linear growth during first three

years of life correlated positively with microbial diversity as well as relative abundance of the genus *Dialister*, among other taxa. An earlier study found that malnourished Bangladeshi children (with weight-for-height Z-scores below -3) harbored immature gut microbiota[85]. Another case-control comparison of Indian children with stunted vs. normal growth found differences in their gut microbiotas[86]. In Europe, a study found associations between the early gut microbiota at the age of three months and BMI at age 5-6 in children from Finland and the Netherlands. These differences were stronger among children with a history of antibiotic use[87]. Indeed, early antibiotic use has been associated with growth in livestock, animal models and humans, reviewed in[88]; an effect which is likely mediated, at least partially, through the gut microbiome[89]. Our results support the hypothesis that the early gut microbiome is an important factor in healthy growth in infancy and early childhood.

The SNP and gene content profiling offer two complementary means of tracking microbial strains in metagenomic data. The SNP based methods usually operate within a small fraction of a genome that serves as a marker region for evaluating evolutionary distance within the population in question[23,52,90]. Evaluating the gene content of microbial strains offers more direct means for functional interpretation of any observed differences[19,20]. We found that in most species these two approaches provide highly concortand phylogenetic population structure, as evidenced by high correlation between SNP haplotype and gene content similarities. However, in some species, such as *F. prausnitzii* and *B. dorei*, these two measures did not correlate. *F. prausnitzii* is a phylogenetically diverse clade, consisting of distinct subspecies potentially confounding methodologies for tracking strains[53]. In the case of *B. dorei,* we isolated and sequenced 8 high quality genomes that confirmed this observation. For these and other similar species, such as *B. adolescentis* and *R. intestinalis*, the observed lack of correlation may stem from the difference between the time-scales at which these two measures operate: rapid adaptation in gene content driven by promiscuous lateral gene transfer (LGT) and gene loss contrasted by a slower, long-term imprint at the SNP level may confound the connection between these two measures[91,92]. The consequence of most of these adaptations for the strain fitness or its symbiosis with the host early in the life, especially in the light of the known immunomodulatory effects that specific strains can have on the human system remains to be elucidated.

The members of the genus *Bacteroides* are highly versatile carbohydrate-utilizers and as such typically represent a large proportion of healthy gut microbiome throughout life[93]. Our analysis revealed that members of this genus harbor some of the largest, highly strain-specific accessory genomes often with hundreds of unique genes per strain. This is mirrored by their ability to adapt the carbohydrate-active enzyme repertoire according to the available resources, modified by diet. Bacteroides targeting prophages are among the most common members of the yet-largely-unexplored human gut virome, providing a plausible mechanism for extensive LGT and genomic plasticity in this genus[94]. Indeed, it has been demonstrated that phages enable LGT between *Staphylococcus aureus* strains[95] and within Enterobacteriaceae family[96]. Supporting this line of thought, a previous investigation of viral contigs in a subset of the DIABIMMUNE stool samples found co-occurrence between multiple viral contigs and *Bacteroides* spp.[97]. Similarly, the most abundant members of the human gut virome, crAss-like phages, were

recently associated with bacteria from the phylum Bacteroidetes, especially *Bacteroides* spp.[98]. Alternatively, highly conserved SNP haplotypes in this genus, implicating long highly-conserved genomic regions, provide another speculative mechanism for LGT: free-floating genetic elements sharing such conserved DNA regions with the recipient genome can be readily transformed in the recipient genome by spontaneous DNA recombination.

This study also contributes several observations on another group of bacteria common in early childhood, *Bifidobacterium*. Within *Bifidobacterium*, there are important differences in HMO processing capabilities even within species making strain level identification of Bifidobacteria crucial. We showed that a *B. longum subsp infantis* can be detected in metagenomic data by both its HMO processing genes and SNP haplotype profiles. A probiotic trial using *B. infantis* as an additive in breast milk during the first weeks of life observed persistent *B. infantis* engraftment and beneficial alterations in intestinal fermentation[99]. Our data corroborates the notion that intracellular HMO utilization in *B. infantis* provides a competitive advantage over other HMO-consuming species, allowing *B. infantis* to dominate the infant gut during breastfeeding.

We observed virtually identical *B. bifidum* strains in 79 mostly Russian stool samples. This analysis demonstrates that microbial strains can be shared on population level and such strain-level trends can be detected from metagenomic data. The observed strain, *B. bifidum* 791, has been patented for medical use in Russia and local regulation allows adding such bacterial components to infant formulas. Indeed, our communication with locals confirmed that this strain is a common component in baby formulas and other infant food products. Therefore, it is plausible that these 34 Russians obtained this strain, which seem to achieve stable engraftment, as a probiotic supplement (in either infant formula or elsewhere). This observation supports the idea that early gut microbial assembly can be intervened by probiotic supplementation[99], which in turn can have beneficial effects such as restoration of healthy growth[85] and protection against immune-mediated diseases[100] or adverse effects of antibiotic courses[101].

Methods

The DIABIMMUNE cohort recruitment took place between September 2008 and July 2011 in Espoo / Finland, Tartu / Estonia and Petrozavodsk / Russia. Families with a newborn with HLA DR-DQ alleles conferring increased risk for autoimmunity, determined by a cord blood test, were invited to join the study. The parents and/or study subjects gave their written informed consent prior to sample collection. The study participants were monitored for infections, use of antibiotics, breastfeeding, introduction of complementary foods, and other life events on study visits at months 3, 6, 12, 18, 24, and 36 from birth. Maternal information and events during the pregnancy were collected using a questionnaire on these visits. Serum samples were collected from all subjects during visits to the clinic at the following time points: 0 (cord blood), 3, 6, 12, 18, 24, and 36 months. Diabetes-associated autoantibodies were analyzed as previously described[40]. The DIABIMMUNE study was conducted according to the guidelines laid down in the Declaration of Helsinki, and all procedures involving human subjects were approved by the local ethical committees of the participating hospitals. More information about the cohort and data collection can be found in other DIABIMMUNE publications[40,41,43] and online at http://www.diabimmune.org/ and https://pubs.broadinstitute.org/diabimmune/.

For statistical association testing described below, the additional information (external variables) of subjects was preprocessed as follows. The external variables were categorized into two categories: generic and complex variables. Here, generic variables' information was available for all subjects, i.e. contained no missing values (maternal age at delivery, gestational diabetes, gestational age days, mode of delivery, gender, country of birth, cohort, and HLA risk class). Complex variables, on the other hand, contained missing values and in many cases required pre-processing and exact defining beforehand (for e.g. antibiotics courses, maternal illnesses during pregnancy, family location when the child was born (urban/rural), daycare attendance, elder siblings, etc.). As breastfeeding information was not available for all the subjects and reduced the sample sizes significantly in cross-sectional analyses, it was not considered a generic variable. The full lists of generic and complex variables can be found in Table ###. While the associations between the generic variables and the gut microbial communities were modeled altogether in one analysis, the associations of complex variables were determined by modeling them one-by-one with all generic variables.

16S rRNA gene sequencing was conducted essentially as previously described in[102]. Paired-end sequencing reads were demultiplexed using ea-utils command line tools (https://code.google.com/p/ea-utils/), and clustered into operational taxonomic units (OTUs) using the UPARSE pipeline[103]. Reads were quality-filtered using the UPARSE quality-filtering threshold of Emax=1, at which the most probable number of base errors per read is zero for filtered reads[104]. Filtered reads were trimmed to a fixed length, singletons removed, and clustered *de novo* into OTUs, with simultaneous chimera filtering. Taxonomic classification of OTUs was performed against the Greengenes version 13.8 16S rDNA database[105]. The full OTU table was filtered by removing samples with less than 3,000 OTU counts, and by removing OTUs appearing in less than 5 % of samples (178 samples). This resulted in an OTU table consisting of 3,204 samples from 289 subjects and 920 OTUs.

PERMANOVA analysis between the external variables and gut microbiomes were performed on 16S rRNA amplicon sequencing data of samples collected roughly at 2 (between 0 and 90 days), 6 (170 and 260 days) and 18 months (510 to 600 days) of age using adonis function in *vegan* R package (default parameters). Per each subject, the sample closest to the exact cross-section time under analysis was chosen, resulting in 140, 184 and 202 samples per time window, respectively. The order of external variables in PERMANOVA model formula was determined by first analyzing each variable individually and then ordering the variables based on the significance of their association (i.e. permuted p-value) from the most significant to the least. Statistical significance of PERMANOVA results was evaluated by permutation test with 10,000 permutations.

Individual associations between bacterial genera and external variables were tested using MaAsLin, which conducts outlier removal, feature selection and linear modeling[106]. Association analyses were performed in both cross-sectional and longitudinal manner. The cross-sectional analyses were conducted on the same samples from the time-windows chosen for the PERMANOVA analyses, where all variables of the analyses (only generic variables or generic and one added complex variable) were used as fixed effects. In the longitudinal analyses, subject IDs were used as a random effect and all the generic variables as well as breastfeeding information were used as fixed effects. In case a complex variable was added to the analysis, it was also used as a fixed effect. With these effect settings in longitudinal analyses, altogether 2586 samples from 237 subjects were available, where the numbers varied according to the complex variable added to the analysis and the amount of missing values it introduced. For both the cross-sectional and longitudinal analyses, genus-level 16S rRNA microbiome data was used for identifying taxonomic level associations of the external variables.

The metagenomic shotgun sequencing was conducted as previously described[40,41,43]. The quality control for the metagenomic shotgun sequencing data was conducted using kneadData v0.4.6.1 with additional automatic adapter detection and trimming at a minimum overlap of 5bp by Trim Galore!. Taxonomic profiles were generated using MetaPhlAn v2.6[107] and functional profiling was done by HUMAnN2 v0.10.0 which provides gene family level (here 90 % similarity) quantifications of microbial genes which are further stratified by contributing organisms. The gene families were further mapped to Gene Ontology (GO) terms as previously described in[43]. Strain SNP haplotypes were generated using StrainPhlAn[23] by requiring minimum coverage of 10 bases for SNP calling ("--min_read_depth 10" command line parameter for sample2markers.py).

Metagenomic reads were assembled into contigs using MegaHIT[108] individually for each sample, followed by an open reading frame prediction using Prodigal[109]. Non-redundant gene catalogue was constructed in a fashion similar to earlier approaches[110] by clustering genes based on sequence similarity at 95 % identity and 90 % coverage of the shorter sequence using CD-HIT[111]. Subsequently, the gene catalogue was merged with IGC gene catalogue[112] using the same criteria to create a more comprehensive reference gene catalogue for the gut microbiome. Only genes detected in Diabimmune samples (~6M) were used in the downstream analysis.

Gene abundance was estimated by mapping quality trimmed reads from each sample to the gene catalogue with BWA[113] which served as an input to binning genes into metagenomic species using canopy clustering[67]. Pangenome for each metagenomic species with at least 400 genes was created by recruiting accessory genes, i.e. genes co-assembled on the same contigs as core genes, i.e. genes binned into metagenomic species, as long as the abundance of the accessory genes was between 10th and 90th percentile of the abundance of core genes in a given sample. Assembled genes were annotated with COG, KEGG and GO terms using eggnog mapper[114] and at species, genus and phylum levels with NCBI RefSeq (version July 2017) as described previously[112].

Phylogenetic trees (**Fig. 3A,C, 5A, S3E,F**) were generated based on StrainPhlAn SNP haplotypes using *phangorn* R package[115]. Briefly, similarities between strain haplotypes were computed using Jukes and Cantor (JC69) model, and an initial tree was constructed using UPGMA hierarchical clustering. The tree was optimized using maximum likelihood method, by iterative optimization of edge lengths, base frequencies and topology. Visualizations were generated using *ggtree* R package. For *Bifidobacterium bifidum* (**Fig. 3C**), strains with >99.5 % sequence similarity were collapsed to a single tip and represented by the strain with the lowest average distance to other strains prior to optimizing the phylogenetic tree.

Bacteroides dorei colonies were isolated from serial dilutions of DIABIMMUNE and PRISM (Prospective Registry in IBD Study at Massachusetts General Hospital) stool samples plated on selective and non selective media after being incubated anaerobically at 37 °C for 72 hours. To isolate high molecular weight DNA for Pacific Biosciences (PacBio) Sequencing, the isolates were grown on brain heart infusion agar supplemented (sBHI) with 10 % fetal bovine serum (Hyclone), 1 % hemin/vitamin K solution (BD), 1 % trace vitamins (ATCC), 1 % trace minerals (ATCC), 0.5 g/L cysteine hydrochloride (Sigma), 1 g/L maltose, 1 g/L fructose (VWR), and 1 g/L cellubiose (Sigma) anaerobically at 37 °C for 72 hours. Colonies were transferred to 30 mL sBHI broth and grown anaerobically for 48 hours. Cells were centrifuged at 4,450 rpm for 10 minutes and supernatant was discarded. DNA was extracted using the Genomic-tip 500/G kit (Qiagen), according to the manufacturer's instruction. After isopropanol treatment, precipitated DNA was spooled and transferred to 70 % ethanol 1.2 mL tube and left to dry in a clean PCR hood for 4 hours. Dried DNA was resuspended in elution buffer (Qiagen). DNA fragment size was measured with 4200 TapeStation (Agilent) using a Genomic DNA ScreenTape (Agilent) prior to sequencing using the PacBio sequencing platform.

PacBio sequencing data of *B. dorei* isolates was assembled into genomes using Celera assembler and Quiver in SMRT Analysis software (PacBio). The assembled *B. dorei* genomes were analyzed using PanPhlAn[24] (default settings) together with five existing isolate genomes in NCBI. The resulting non-redundant gene catalogue was annotated by translated DIAMOND search[116] against the UniRef90 and UniRef50 databases, and by enforcing UniRef's clustering criteria. We primarily used UniRef90 annotations, if available, but applied UniRef50 annotation in absence of UniRef90 annotation. *Ruminococcus gnavus* pangenome analysis was conducted using an existing PanPhlAn-generated pangenome as previously described[16], which was annotated using UniRef databases as described above for the *B. dorei* pangenome.

B. longum HMO gene presence in the metagenomic samples (**Fig. 3A**) was determined as follows. We identified UniRef90 gene families corresponding to the protein sequences in *B. infantis* HMO gene cluster (protein sequences Blon_2331-Blon_2361[71] in NCBI protein sequence database) by translated BLAST search against *B. longum* pangenome in ChocoPhlAn pangenome collection[66] utilized by HUMAnN2. Specifically, we required >=90 % alignment identify and >=80 % mutual coverage (corresponding to the definition of UniRef90 gene families) and accepted only the best hit per protein sequence. Combining this information with HUMAnN2 species-stratified UniRef90 gene family quantification enabled calling these genes present given that they had sufficient read coverage, here defined as log10(counts-per-million / *B. longum* relative abundance) > 1.

Contributional diversities of the metagenomic functions were analyzed as previously described in [Franzosa et al., under review]. Briefly, stratified abundances of metagenomic functions were first renormalized after excluding any "unclassified" relative abundance. Contributional diversity for a given metagenomic function was then calculated by applying ecological similarity measures to the stratified abundance of that function; Gini-Simpson index was used for alpha-diversity and Bray-Curtis dissimilarity was used for beta-diversity.

References

1 Kundu, P., Blacher, E., Elinav, E. & Pettersson, S. Our Gut Microbiome: The Evolving Inner Self. *Cell* **171**, 1481-1493, doi:10.1016/j.cell.2017.11.024 (2017).

2 Rodriguez, J. M. *et al.* The composition of the gut microbiota throughout life, with an emphasis on early life. *Microb Ecol Health Dis* **26**, 26050, doi:10.3402/mehd.v26.26050 (2015).

3 Rewers, M. & Ludvigsson, J. Environmental risk factors for type 1 diabetes. *Lancet* **387**, 2340-2348, doi:10.1016/S0140-6736(16)30507-4 (2016).

4 Knip, M. & Siljander, H. The role of the intestinal microbiota in type 1 diabetes mellitus. *Nat Rev Endocrinol* **12**, 154-167, doi:10.1038/nrendo.2015.218 (2016).

5 Arrieta, M. C. *et al.* Early infancy microbial and metabolic alterations affect risk of childhood asthma. *Sci Transl Med* **7**, 307ra152, doi:10.1126/scitranslmed.aab2271 (2015).

6 Abrahamsson, T. R. *et al.* Low gut microbiota diversity in early infancy precedes asthma at school age. *Clin Exp Allergy* **44**, 842-850, doi:10.1111/cea.12253 (2014).

7 Arvonen, M. *et al.* Gut microbiota-host interactions and juvenile idiopathic arthritis. *Pediatr Rheumatol Online J* **14**, 44, doi:10.1186/s12969-016-0104-6 (2016).

8 Simonyte Sjodin, K., Vidman, L., Ryden, P. & West, C. E. Emerging evidence of the role of gut microbiota in the development of allergic diseases. *Curr Opin Allergy Clin Immunol* **16**, 390-395, doi:10.1097/ACI.0000000000000277 (2016).

9 Lewis, J. D. *et al.* Inflammation, Antibiotics, and Diet as Environmental Stressors of the Gut Microbiome in Pediatric Crohn's Disease. *Cell Host Microbe* **18**, 489-500, doi:10.1016/j.chom.2015.09.008 (2015).

10 Bach, J. F. The hygiene hypothesis in autoimmunity: the role of pathogens and commensals. *Nat Rev Immunol* **18**, 105-120, doi:10.1038/nri.2017.111 (2018).

11 Haahtela, T. *et al.* The biodiversity hypothesis and allergic disease: world allergy organization position statement. *World Allergy Organ J* **6**, 3, doi:10.1186/1939-4551-6-3 (2013).

12 Thaiss, C. A., Zmora, N., Levy, M. & Elinav, E. The microbiome and innate immunity. *Nature* **535**, 65-74, doi:10.1038/nature18847 (2016).
13 Honda, K. & Littman, D. R. The microbiota in adaptive immune homeostasis and disease. *Nature* **535**, 75-84, doi:10.1038/nature18848 (2016).
14 Arthur, J. C. *et al.* Intestinal inflammation targets cancer-inducing activity of the microbiota. *Science* **338**, 120-123, doi:10.1126/science.1224820 (2012).
15 Lebreton, F. *et al.* Emergence of epidemic multidrug-resistant Enterococcus faecium from animal and commensal strains. *MBio* **4**, doi:10.1128/mBio.00534-13 (2013).
16 Hall, A. B. *et al.* A novel Ruminococcus gnavus clade enriched in inflammatory bowel disease patients. *Genome Med* **9**, 103, doi:10.1186/s13073-017-0490-5 (2017).
17 Schonherr-Hellec, S. *et al.* Clostridial strain-specific characteristics associated with necrotizing enterocolitis. *Appl Environ Microbiol*, doi:10.1128/AEM.02428-17 (2018).
18 Bron, P. A., van Baarlen, P. & Kleerebezem, M. Emerging molecular insights into the interaction between probiotics and the host intestinal mucosa. *Nat Rev Microbiol* **10**, 66-78, doi:10.1038/nrmicro2690 (2011).
19 Ward, D. V. *et al.* Metagenomic Sequencing with Strain-Level Resolution Implicates Uropathogenic E. coli in Necrotizing Enterocolitis and Mortality in Preterm Infants. *Cell Rep* **14**, 2912-2924, doi:10.1016/j.celrep.2016.03.015 (2016).
20 Hazen, T. H. *et al.* Genomic diversity of EPEC associated with clinical presentations of differing severity. *Nat Microbiol* **1**, 15014, doi:10.1038/nmicrobiol.2015.14 (2016).
21 Sela, U., Euler, C. W., Correa da Rosa, J. & Fischetti, V. A. Strains of bacterial species induce a greatly varied acute adaptive immune response: The contribution of the accessory genome. *PLoS Pathog* **14**, e1006726, doi:10.1371/journal.ppat.1006726 (2018).
22 Luo, C. *et al.* ConStrains identifies microbial strains in metagenomic datasets. *Nat Biotechnol* **33**, 1045-1052, doi:10.1038/nbt.3319 (2015).
23 Truong, D. T., Tett, A., Pasolli, E., Huttenhower, C. & Segata, N. Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Res* **27**, 626-638, doi:10.1101/gr.216242.116 (2017).
24 Scholz, M. *et al.* Strain-level microbial epidemiology and population genomics from shotgun metagenomics. *Nat Methods* **13**, 435-438, doi:10.1038/nmeth.3802 (2016).
25 Perez-Munoz, M. E., Arrieta, M. C., Ramer-Tait, A. E. & Walter, J. A critical assessment of the "sterile womb" and "in utero colonization" hypotheses: implications for research on the pioneer infant microbiome. *Microbiome* **5**, 48, doi:10.1186/s40168-017-0268-4 (2017).
26 Dominguez-Bello, M. G. *et al.* Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proc Natl Acad Sci U S A* **107**, 11971-11975, doi:10.1073/pnas.1002601107 (2010).
27 Tamburini, S., Shen, N., Wu, H. C. & Clemente, J. C. The microbiome in early life: implications for health outcomes. *Nat Med* **22**, 713-722, doi:10.1038/nm.4142 (2016).
28 McGuire, M. K. & McGuire, M. A. Human milk: mother nature's prototypical probiotic food? *Adv Nutr* **6**, 112-123, doi:10.3945/an.114.007435 (2015).
29 Jost, T., Lacroix, C., Braegger, C. & Chassard, C. Impact of human milk bacteria and oligosaccharides on neonatal gut microbiota establishment and gut health. *Nutr Rev* **73**, 426-437, doi:10.1093/nutrit/nuu016 (2015).
30 Gomez-Gallego, C., Garcia-Mantrana, I., Salminen, S. & Collado, M. C. The human milk microbiome and factors influencing its composition and activity. *Semin Fetal Neonatal Med* **21**, 400-405, doi:10.1016/j.siny.2016.05.003 (2016).
31 Charbonneau, M. R. *et al.* A microbial perspective of human developmental biology. *Nature* **535**, 48-55, doi:10.1038/nature18845 (2016).

32 de Goffau, M. C. *et al.* Aberrant gut microbiota composition at the onset of type 1 diabetes in young children. *Diabetologia* **57**, 1569-1577, doi:10.1007/s00125-014-3274-0 (2014).

33 de Goffau, M. C. *et al.* Fecal microbiota composition differs between children with beta-cell autoimmunity and those without. *Diabetes* **62**, 1238-1244, doi:10.2337/db12-0526 (2013).

34 Davis-Richardson, A. G. *et al.* Bacteroides dorei dominates gut microbiome prior to autoimmunity in Finnish children at high risk for type 1 diabetes. *Front Microbiol* **5**, 678, doi:10.3389/fmicb.2014.00678 (2014).

35 Endesfelder, D. *et al.* Compromised gut microbiota networks in children with anti-islet cell autoimmunity. *Diabetes* **63**, 2006-2014, doi:10.2337/db13-1676 (2014).

36 Mejia-Leon, M. E., Petrosino, J. F., Ajami, N. J., Dominguez-Bello, M. G. & de la Barca, A. M. Fecal microbiota imbalance in Mexican children with type 1 diabetes. *Sci Rep* **4**, 3814, doi:10.1038/srep03814 (2014).

37 Alkanani, A. K. *et al.* Alterations in Intestinal Microbiota Correlate With Susceptibility to Type 1 Diabetes. *Diabetes* **64**, 3510-3520, doi:10.2337/db14-1847 (2015).

38 Endesfelder, D. *et al.* Towards a functional hypothesis relating anti-islet cell autoimmunity to the dietary impact on microbial communities and butyrate production. *Microbiome* **4**, 17, doi:10.1186/s40168-016-0163-4 (2016).

39 Maffeis, C. *et al.* Association between intestinal permeability and faecal microbiota composition in Italian children with beta cell autoimmunity at risk for type 1 diabetes. *Diabetes Metab Res Rev* **32**, 700-709, doi:10.1002/dmrr.2790 (2016).

40 Kostic, A. D. *et al.* The dynamics of the human infant gut microbiome in development and in progression toward type 1 diabetes. *Cell Host Microbe* **17**, 260-273, doi:10.1016/j.chom.2015.01.001 (2015).

41 Yassour, M. *et al.* Natural history of the infant gut microbiome and impact of antibiotic treatment on bacterial strain diversity and stability. *Sci Transl Med* **8**, 343ra381, doi:10.1126/scitranslmed.aad0917 (2016).

42 Kallionpaa, H. *et al.* Standard of hygiene and immune adaptation in newborn infants. *Clin Immunol* **155**, 136-147, doi:10.1016/j.clim.2014.09.009 (2014).

43 Vatanen, T. *et al.* Variation in Microbiome LPS Immunogenicity Contributes to Autoimmunity in Humans. *Cell* **165**, 842-853, doi:10.1016/j.cell.2016.04.007 (2016).

44 Mustonen, N. *et al.* Early childhood infections precede development of beta-cell autoimmunity and type 1 diabetes in children with HLA-conferred disease risk. *Pediatr Diabetes* **19**, 293-299, doi:10.1111/pedi.12547 (2018).

45 Simre, K. *et al.* Exploring the risk factors for differences in the cumulative incidence of coeliac disease in two neighboring countries: the prospective DIABIMMUNE study. *Dig Liver Dis* **48**, 1296-1301, doi:10.1016/j.dld.2016.06.029 (2016).

46 Ruokolainen, L. *et al.* Green areas around homes reduce atopic sensitization in children. *Allergy* **70**, 195-202, doi:10.1111/all.12545 (2015).

47 Reinert-Hartwall, L. *et al.* Th1/Th17 plasticity is a marker of advanced beta cell autoimmunity and impaired glucose tolerance in humans. *J Immunol* **194**, 68-75, doi:10.4049/jimmunol.1401653 (2015).

48 Ege, M. J. *et al.* Exposure to environmental microorganisms and childhood asthma. *N Engl J Med* **364**, 701-709, doi:10.1056/NEJMoa1007302 (2011).

49 Lax, S. *et al.* Longitudinal analysis of microbial interaction between humans and the indoor environment. *Science* **345**, 1048-1052, doi:10.1126/science.1254529 (2014).

50 Yatsunenko, T. *et al.* Human gut microbiome viewed across age and geography. *Nature* **486**, 222-227, doi:10.1038/nature11053 (2012).

51 Bokulich, N. A. *et al.* Antibiotics, birth mode, and diet shape microbiome maturation during early life. *Sci Transl Med* **8**, 343ra382, doi:10.1126/scitranslmed.aad7121 (2016).

52 Lloyd-Price, J. *et al.* Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature* **550**, 61-66, doi:10.1038/nature23889 (2017).

53 He, Q. *et al.* Two distinct metacommunities characterize the gut microbiota in Crohn's disease patients. *Gigascience* **6**, 1-11, doi:10.1093/gigascience/gix050 (2017).

54 Atarashi, K. *et al.* Ectopic colonization of oral bacteria in the intestine drives TH1 cell induction and inflammation. *Science* **358**, 359-365, doi:10.1126/science.aan4526 (2017).

55 Scher, J. U. *et al.* Expansion of intestinal Prevotella copri correlates with enhanced susceptibility to arthritis. *Elife* **2**, e01202, doi:10.7554/eLife.01202 (2013).

56 Browne, H. P. *et al.* Culturing of 'unculturable' human microbiota reveals novel taxa and extensive sporulation. *Nature* **533**, 543-546, doi:10.1038/nature17645 (2016).

57 Lagier, J. C. *et al.* Culture of previously uncultured members of the human gut microbiota by culturomics. *Nat Microbiol* **1**, 16203, doi:10.1038/nmicrobiol.2016.203 (2016).

58 Schloissnig, S. *et al.* Genomic variation landscape of the human gut microbiome. *Nature* **493**, 45-50, doi:10.1038/nature11711 (2013).

59 Franzosa, E. A. *et al.* Identifying personal microbiomes using metagenomic codes. *Proc Natl Acad Sci U S A* **112**, E2930-2938, doi:10.1073/pnas.1423854112 (2015).

60 Lange, A. *et al.* Extensive Mobilome-Driven Genome Diversification in Mouse Gut-Associated Bacteroides vulgatus mpk. *Genome Biol Evol* **8**, 1197-1207, doi:10.1093/gbe/evw070 (2016).

61 de la Cruz, F. & Davies, J. Horizontal gene transfer and the origin of species: lessons from bacteria. *Trends Microbiol* **8**, 128-133 (2000).

62 McInerney, J. O., McNally, A. & O'Connell, M. J. Why prokaryotes have pangenomes. *Nat Microbiol* **2**, 17040, doi:10.1038/nmicrobiol.2017.40 (2017).

63 Brito, I. L. *et al.* Mobile genes in the human microbiome are structured from global to individual scales. *Nature* **535**, 435-439, doi:10.1038/nature18927 (2016).

64 Smillie, C. S. *et al.* Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* **480**, 241-244, doi:10.1038/nature10571 (2011).

65 Rothschild, D. *et al.* Environment dominates over host genetics in shaping human gut microbiota. *Nature* **555**, 210-215, doi:10.1038/nature25973 (2018).

66 Huang, K. *et al.* MetaRef: a pan-genomic database for comparative and community microbial genomics. *Nucleic Acids Res* **42**, D617-624, doi:10.1093/nar/gkt1078 (2014).

67 Nielsen, H. B. *et al.* Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat Biotechnol* **32**, 822-828, doi:10.1038/nbt.2939 (2014).

68 Land, M. *et al.* Insights from 20 years of bacterial genome sequencing. *Funct Integr Genomics* **15**, 141-161, doi:10.1007/s10142-015-0433-4 (2015).

69 Bottacini, F., van Sinderen, D. & Ventura, M. Omics of bifidobacteria: research and insights into their health-promoting activities. *Biochem J* **474**, 4137-4152, doi:10.1042/BCJ20160756 (2017).

70 Sela, D. A. & Mills, D. A. Nursing our microbiota: molecular linkages between bifidobacteria and milk oligosaccharides. *Trends Microbiol* **18**, 298-307, doi:10.1016/j.tim.2010.03.008 (2010).

71 Sela, D. A. *et al.* The genome sequence of Bifidobacterium longum subsp. infantis reveals adaptations for milk utilization within the infant microbiome. *Proc Natl Acad Sci U S A* **105**, 18964-18969, doi:10.1073/pnas.0809584105 (2008).

72 Garrido, D. *et al.* A novel gene cluster allows preferential utilization of fucosylated milk oligosaccharides in Bifidobacterium longum subsp. longum SC596. *Sci Rep* **6**, 35045, doi:10.1038/srep35045 (2016).

73 Sela, D. A. Bifidobacterial utilization of human milk oligosaccharides. *Int J Food Microbiol* **149**, 58-64, doi:10.1016/j.ijfoodmicro.2011.01.025 (2011).

74 van den Bogert, B. *et al.* Diversity of human small intestinal Streptococcus and Veillonella populations. *FEMS Microbiol Ecol* **85**, 376-388, doi:10.1111/1574-6941.12127 (2013).
75 Buffie, C. G. & Pamer, E. G. Microbiota-mediated colonization resistance against intestinal pathogens. *Nat Rev Immunol* **13**, 790-801, doi:10.1038/nri3535 (2013).
76 Wang, M. *et al.* Fecal microbiota composition of breast-fed infants is correlated with human milk oligosaccharides consumed. *J Pediatr Gastroenterol Nutr* **60**, 825-833, doi:10.1097/MPG.0000000000000752 (2015).
77 Guaraldi, F. & Salvatori, G. Effect of breast and formula feeding on gut microbiota shaping in newborns. *Front Cell Infect Microbiol* **2**, 94, doi:10.3389/fcimb.2012.00094 (2012).
78 Hille, F. *et al.* The Biology of CRISPR-Cas: Backward and Forward. *Cell* **172**, 1239-1259, doi:10.1016/j.cell.2017.11.032 (2018).
79 Doron, S. *et al.* Systematic discovery of antiphage defense systems in the microbial pangenome. *Science* **359**, doi:10.1126/science.aar4120 (2018).
80 Rho, M., Wu, Y. W., Tang, H., Doak, T. G. & Ye, Y. Diverse CRISPRs evolving in human microbiomes. *PLoS Genet* **8**, e1002441, doi:10.1371/journal.pgen.1002441 (2012).
81 Grissa, I., Vergnaud, G. & Pourcel, C. The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics* **8**, 172, doi:10.1186/1471-2105-8-172 (2007).
82 Abubucker, S. *et al.* Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Comput Biol* **8**, e1002358, doi:10.1371/journal.pcbi.1002358 (2012).
83 Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-29, doi:10.1038/75556 (2000).
84 Morris, J. J., Lenski, R. E. & Zinser, E. R. The Black Queen Hypothesis: evolution of dependencies through adaptive gene loss. *MBio* **3**, doi:10.1128/mBio.00036-12 (2012).
85 Subramanian, S. *et al.* Persistent gut microbiota immaturity in malnourished Bangladeshi children. *Nature* **510**, 417-421, doi:10.1038/nature13421 (2014).
86 Dinh, D. M. *et al.* Longitudinal Analysis of the Intestinal Microbiota in Persistently Stunted Young Children in South India. *PLoS One* **11**, e0155405, doi:10.1371/journal.pone.0155405 (2016).
87 Korpela, K. *et al.* Childhood BMI in relation to microbiota in infancy and lifetime antibiotic use. *Microbiome* **5**, 26, doi:10.1186/s40168-017-0245-y (2017).
88 Cox, L. M. & Blaser, M. J. Antibiotics in early life and obesity. *Nat Rev Endocrinol* **11**, 182-190, doi:10.1038/nrendo.2014.210 (2015).
89 Coates, M. E., Fuller, R., Harrison, G. F., Lev, M. & Suffolk, S. F. A comparison of the growth of chicks in the Gustafsson germ-free apparatus and in a conventional environment, with and without dietary supplements of penicillin. *Br J Nutr* **17**, 141-150 (1963).
90 Korpela, K. *et al.* Selective maternal seeding and environment shape the human gut microbiome. *Genome Res* **28**, 561-568, doi:10.1101/gr.233940.117 (2018).
91 Vos, M., Hesselman, M. C., Te Beek, T. A., van Passel, M. W. J. & Eyre-Walker, A. Rates of Lateral Gene Transfer in Prokaryotes: High but Why? *Trends Microbiol* **23**, 598-605, doi:10.1016/j.tim.2015.07.006 (2015).
92 Andreani, N. A., Hesse, E. & Vos, M. Prokaryote genome fluidity is dependent on effective population size. *ISME J* **11**, 1719-1721, doi:10.1038/ismej.2017.36 (2017).
93 El Kaoutari, A., Armougom, F., Gordon, J. I., Raoult, D. & Henrissat, B. The abundance and variety of carbohydrate-active enzymes in the human gut microbiota. *Nat Rev Microbiol* **11**, 497-504, doi:10.1038/nrmicro3050 (2013).

94 Canchaya, C., Fournous, G., Chibani-Chennoufi, S., Dillmann, M. L. & Brussow, H. Phage as agents of lateral gene transfer. *Curr Opin Microbiol* **6**, 417-424 (2003).

95 Haaber, J. *et al.* Bacterial viruses enable their host to acquire antibiotic resistance genes from neighbouring cells. *Nat Commun* **7**, 13333, doi:10.1038/ncomms13333 (2016).

96 Colavecchio, A., Cadieux, B., Lo, A. & Goodridge, L. D. Bacteriophages Contribute to the Spread of Antibiotic Resistance Genes among Foodborne Pathogens of the Enterobacteriaceae Family - A Review. *Front Microbiol* **8**, 1108, doi:10.3389/fmicb.2017.01108 (2017).

97 Zhao, G. *et al.* Intestinal virome changes precede autoimmunity in type I diabetes-susceptible children. *Proc Natl Acad Sci U S A* **114**, E6166-E6175, doi:10.1073/pnas.1706359114 (2017).

98 Yutin, N. *et al.* Discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut. *Nat Microbiol* **3**, 38-46, doi:10.1038/s41564-017-0053-y (2018).

99 Frese, S. A. *et al.* Persistence of Supplemented Bifidobacterium longum subsp. infantis EVC001 in Breastfed Infants. *mSphere* **2**, doi:10.1128/mSphere.00501-17 (2017).

100 Uusitalo, U. *et al.* Association of Early Exposure of Probiotics and Islet Autoimmunity in the TEDDY Study. *JAMA Pediatr* **170**, 20-28, doi:10.1001/jamapediatrics.2015.2757 (2016).

101 Fox, M. J., Ahuja, K. D., Robertson, I. K., Ball, M. J. & Eri, R. D. Can probiotic yogurt prevent diarrhoea in children on antibiotics? A double-blind, randomised, placebo-controlled study. *BMJ Open* **5**, e006474, doi:10.1136/bmjopen-2014-006474 (2015).

102 Gevers, D. *et al.* The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host Microbe* **15**, 382-392, doi:10.1016/j.chom.2014.02.005 (2014).

103 Edgar, R. C. UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat Methods* **10**, 996-998, doi:10.1038/nmeth.2604 (2013).

104 Edgar, R. C. & Flyvbjerg, H. Error filtering, pair assembly and error correction for next-generation sequencing reads. *Bioinformatics* **31**, 3476-3482, doi:10.1093/bioinformatics/btv401 (2015).

105 McDonald, D. *et al.* An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J* **6**, 610-618, doi:10.1038/ismej.2011.139 (2012).

106 Morgan, X. C. *et al.* Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol* **13**, R79, doi:10.1186/gb-2012-13-9-r79 (2012).

107 Segata, N. *et al.* Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat Methods* **9**, 811-814, doi:10.1038/nmeth.2066 (2012).

108 Li, D., Liu, C. M., Luo, R., Sadakane, K. & Lam, T. W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674-1676, doi:10.1093/bioinformatics/btv033 (2015).

109 Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119, doi:10.1186/1471-2105-11-119 (2010).

110 Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59-65, doi:10.1038/nature08821 (2010).

111 Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150-3152, doi:10.1093/bioinformatics/bts565 (2012).

112 Li, J. *et al.* An integrated catalog of reference genes in the human gut microbiome. *Nat Biotechnol* **32**, 834-841, doi:10.1038/nbt.2942 (2014).

113 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

114 Huerta-Cepas, J. *et al.* Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Mol Biol Evol* **34**, 2115-2122, doi:10.1093/molbev/msx148 (2017).
115 Schliep, K. P. phangorn: phylogenetic analysis in R. *Bioinformatics* **27**, 592-593, doi:10.1093/bioinformatics/btq706 (2011).
116 Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**, 59-60, doi:10.1038/nmeth.3176 (2015).

Supplementary Figures

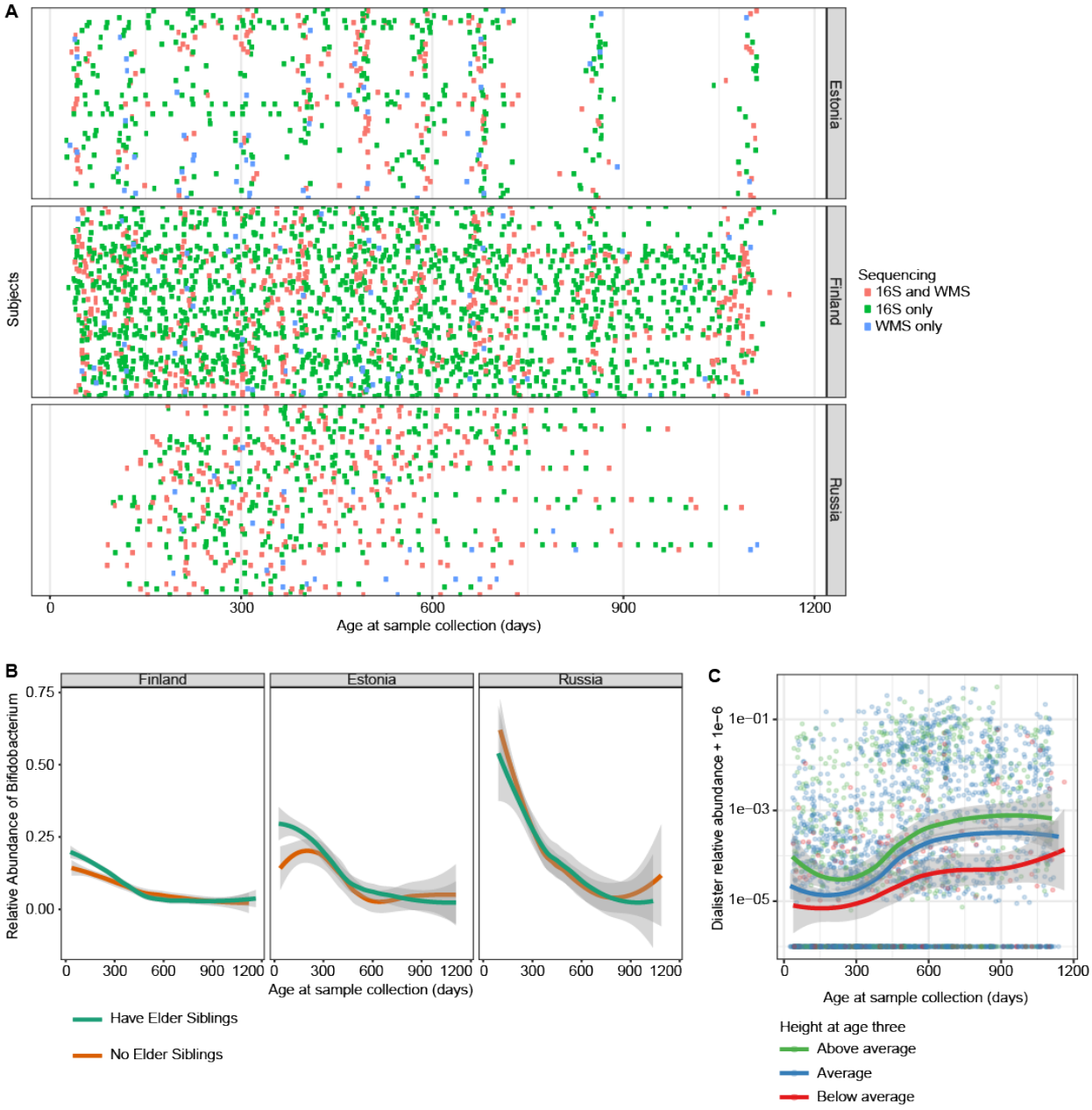

Figure S1. A Samples analyzed by 16S rRNA amplicon and metagenomic sequencing. Rows represent subjects. **B** Average relative abundance of *Bifidobacterium* spp. in 16S sequencing profiles longitudinally stratified by country and presence of elder siblings. The curves show locally weighted scatterplot smoothing (LOESS) for the relative abundances and shaded area shows 95 % confidence interval for each fit, as implemented in geom_smooth() function in ggplot2 R package. **C** Mean relative abundance of *Dialister* spp. In 16S sequencing profiles longitudinally stratified by subjects' height at age three. Weight categories were defined as follows; above average: weight z-score > 1, average: -1 < weight z-score ≤ 1, below average: weight z-score ≤ -1.

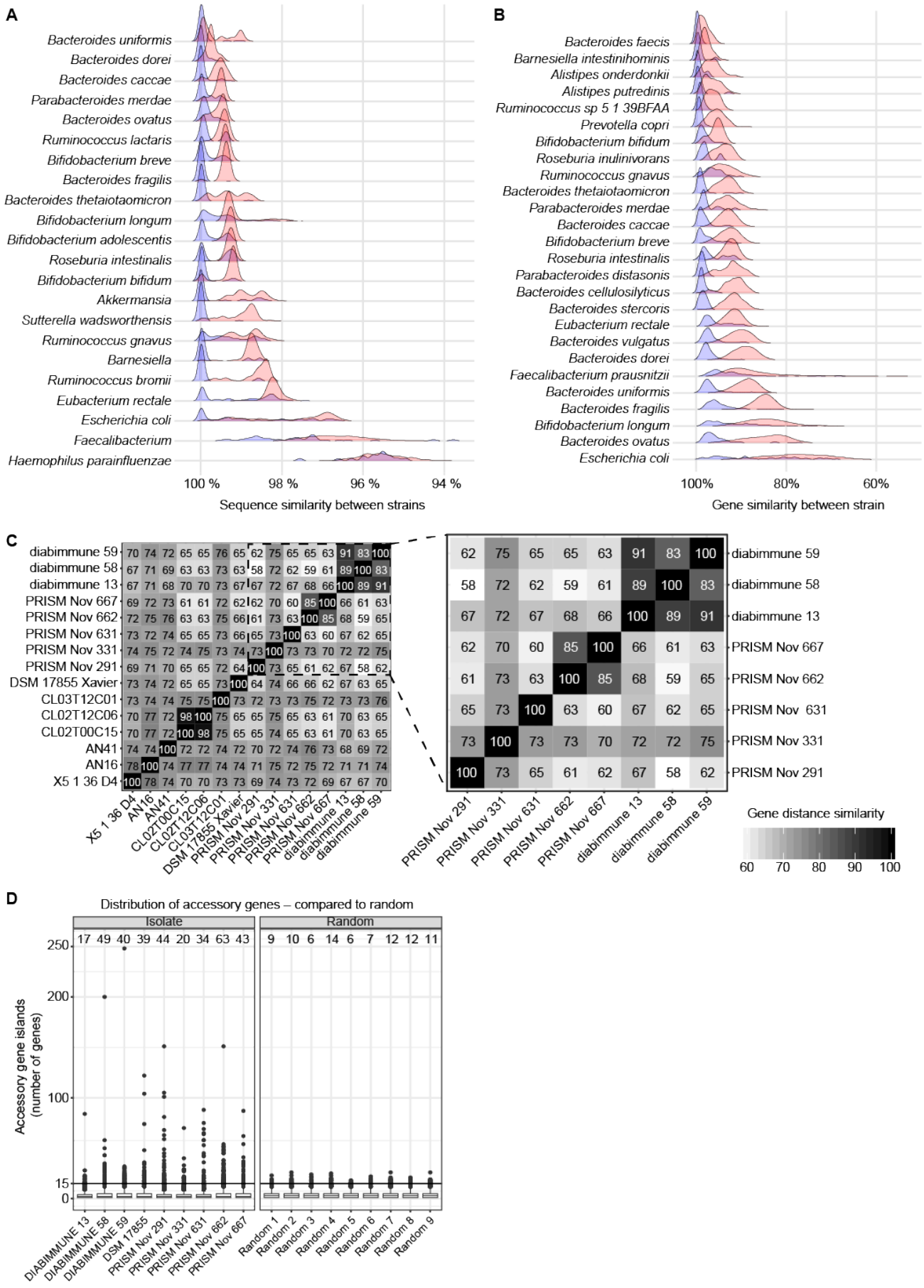
A
Bacteroides uniformis
Bacteroides dorei
Bacteroides caccae
Parabacteroides merdae
Bacteroides ovatus
Ruminococcus lactaris
Bifidobacterium breve
Bacteroides fragilis
Bacteroides thetaiotaomicron
Bifidobacterium longum
Bifidobacterium adolescentis
Roseburia intestinalis
Bifidobacterium bifidum
Akkermansia
Sutterella wadsworthensis
Ruminococcus gnavus
Barnesiella
Ruminococcus bromii
Eubacterium rectale
Escherichia coli
Faecalibacterium
Haemophilus parainfluenzae
100 %
98 %
96 %
94 %
Sequence similarity between strains
B
Bacteroides faecis
Barnesiella intestinihominis
Alistipes onderdonkii
Alistipes putredinis
Ruminococcus sp 5 1 39BFAA
Prevotella copri
Bifidobacterium bifidum
Roseburia inulinivorans
Ruminococcus gnavus
Bacteroides thetaiotaomicron
Parabacteroides merdae
Bacteroides caccae
Bifidobacterium breve
Roseburia intestinalis
Parabacteroides distasonis
Bacteroides cellulosilyticus
Bacteroides stercoris
Eubacterium rectale
Bacteroides vulgatus
Bacteroides dorei
Faecalibacterium prausnitzii
Bacteroides uniformis
Bacteroides fragilis
Bifidobacterium longum
Bacteroides ovatus
Escherichia coli
100%
80%
60%
Gene similarity between strain
C
diabimmune 59
diabimmune 58
diabimmune 13
PRISM Nov 667
PRISM Nov 662
PRISM Nov 631
PRISM Nov 331
PRISM Nov 291
DSM 17855 Xavier
CL03T12C01
CL02T12C06
CL02T00C15
AN41
AN16
X5 1 36 D4
Gene distance similarity
60
70
80
90
100
D
Distribution of accessory genes – compared to random
Isolate
Random
Accessory gene islands (number of genes)
250
200
100
15
0
17 49 40 39 44 20 34 63 43
9 10 6 14 6 7 12 12 11
DIABIMMUNE 13
DIABIMMUNE 58
DIABIMMUNE 59
DSM 17855
PRISM Nov 291
PRISM Nov 331
PRISM Nov 631
PRISM Nov 662
PRISM Nov 667
Random 1
Random 2
Random 3
Random 4
Random 5
Random 6
Random 7
Random 8
Random 9

Figure S2. A Density plots of the SNP haplotype similarities per species based on all pairwise comparisons (dominant strain per species per sample) and stratified to intra-subject and inter-subject comparisons; data in Fig. 2A represented as a density plot. **B** Gene content similarities per species, evaluated on HUMAnN2 gene family profiles. Briefly, HUMAnN2 quantifies gene family abundances per species based on pangenomes that were constructed using the NCBI isolate genomes. **C D**

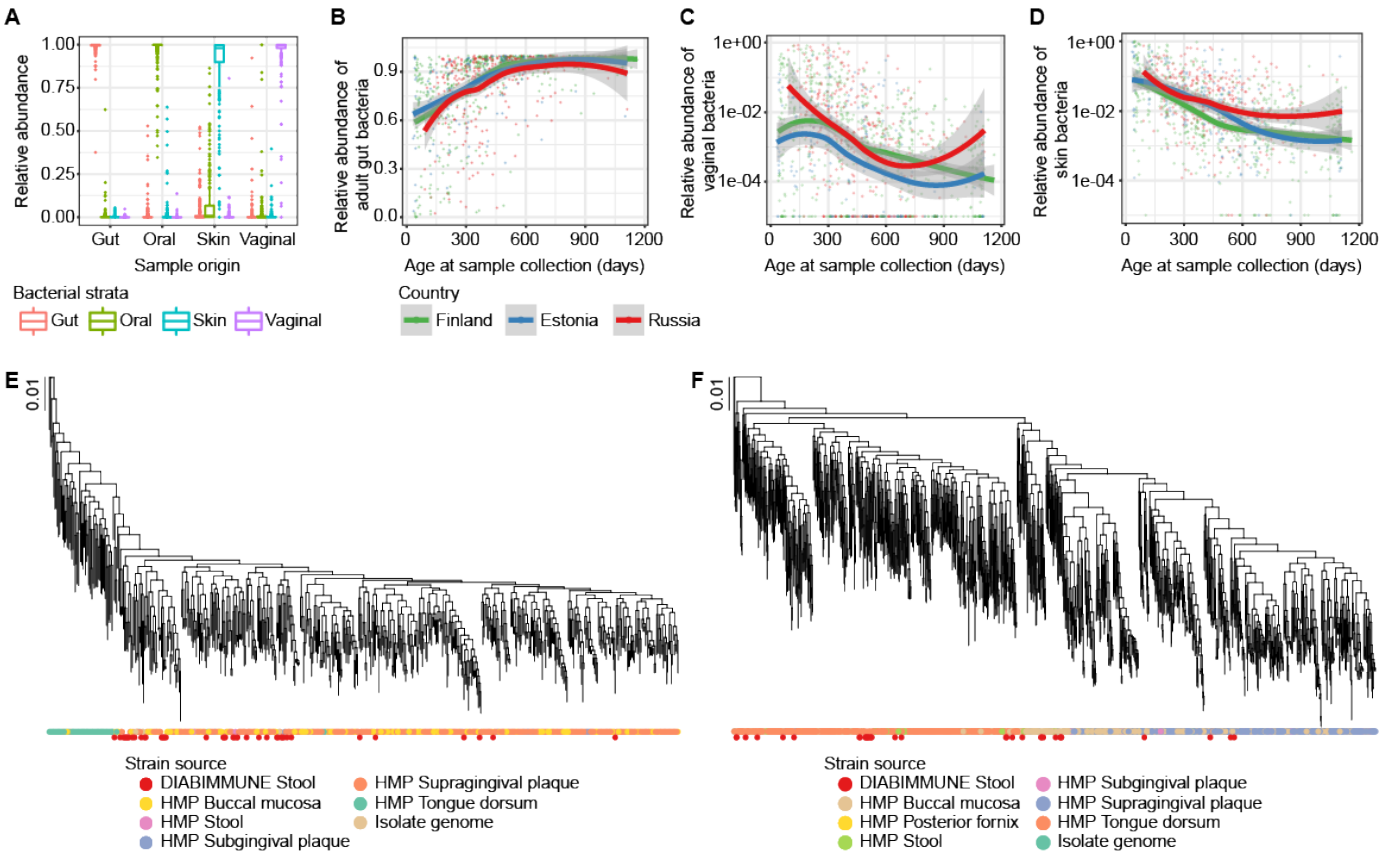

Figure S3. Relations between the adult gut microbiome in HMP and the gut microbiome in DIABIMMUNE. A Boxplot of total relative abundances of bacteria typical of different body areas in HMP data. Each bacterial species in HMP was assigned to one of the four strata by the body area with the highest mean relative abundance of the given species. Color shows the bacterial strata and x-axis shows their total relative abundances in different body areas. See also Table S6. **B-D** Total relative abundance of bacterial species classified as (**B**) gut, (**C**) vaginal and (**D**) skin species in DIABIMMUNE samples longitudinally. The color shows the country of origin and the curves show LOESS fit as detailed in caption for Fig S1. **E-F** Phylogenetic tree of (**E**) *Veillonella parvula* and (**F**) *Haemophilus parainfluenzae* strains in HMP and DIABIMMUNE samples.

Supplementary Table captions

Table S1. PERMANOVA results. Multiple extrinsic and intrinsic factors were analyzed for connections with microbial composition using PERMANOVA. See *PERMANOVA Descriptions* sheet for details.

Table S2. Microbial alpha-diversity. Multiple extrinsic and intrinsic factors were analyzed for connections with microbial alpha-diversity using mixed effects linear modeling. See *Alpha div. Tests Descriptions* sheet for details.

Table S3. Taxonomic associations. Multiple extrinsic and intrinsic factors were analyzed for connections with microbial taxa using MaAsLin linear modeling framework. See *MaAsLin Descriptions* sheet for details.

Table S4. Strain diversity of gut microbial species. Diversity of strains within microbial species were analyzed by SNP haplotyping and gene content on metagenomic assemblies. The table supplements **Fig. 2A-C** with additional statistics. MSA = multiple sequence alignment, *MSA length* gives the effective length of the SNP haplotypes per species.

Table S5. Extended *B. dorei* pangenome. Gene families on extended B. dorei pangenome constructed using seven NCBI isolate genomes and eight additional isolates sequenced in this study. Gene families were annotated using UniRef gene family annotations and presence (1) / absence (0) on each isolate is shown.

Table S6.

Table S7. Bacterial species by body site. Mean relative abundance of bacterial species in HMP data in four body sites in HMP (adult gut, skin, oral cavity, or vagina) and in DIABIMMUNE gut communities. Each species was assigned to a body site given by the highest mean relative abundance in HMP data.

Table S8. CRISPR system genes in DIABIMMUNE metagenomes. The metagenomes were analyzed for CRISPR system genes indicate of speciation; the table lists species specific CRISPR genes with within-species prevalence between 25 % and 75 %, i.e., these genes were carried only a subset of strains in a given species. To obtain confident gene presence calls and avoid false positive hits, only samples where the species in question had relative abundance > 5% were included in the analysis; frequency of such samples is given in column D.

Table S9. *R. gnavus* genes distinguishing subspecies clades. Thousand *R. gnavus* genes with largest difference in prevalence between the subspecies clades 1 and 2 identified by phylogenetic analysis. Annotations are given by translated DIAMOND search against UniRef databases.

Table S10. Contributional diversities of Biological process GO terms. We applied ecological similarity indices (alpha- and beta-diversity) to contributional breakdown (compositional profiles of the species specific contributions to the GO term in question) of 365 biological process GO terms. The tables gives mean and median alpha- and beta-diversities per GO term. For beta-diversities, these measures were further stratified into inter- and intra-subject comparisons. For alpha diversities, we measured Pearson correlation with age and corrected the statistical significance for multiple testing using Benjamini-Hochberg technique.