Supplementary Information for "Genomic variation and strain-specific functional adaptation in the human gut microbiome during early life"

This PDF contains:

Other Supplementary Material includes:

Supplementary Note 1

To extend the understanding of early microbial development in connection with external variables, we analyzed 16S data (n=3,204 samples) using both omnibus and individual association tests. We first compared the taxonomic profiles between 16S amplicon and metagenomic sequencing on the genus level and observed an average Bray-Curtis dissimilarity of 0.283 (standard deviation 0.183). This reflects known technical differences between the techniques arising from, for example, differences between the reference databases and the ability of 16S sequencing to detect taxa with lower relative abundances. 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 courses during pregnancy (permutation test, q-value=0.029) were associated with microbial composition shifts in the earliest stool samples collected at 2 months of age (**Supplementary Table 2**).

We next associated the gut microbial diversities (Chao1 richness, Shannon's diversity index) with the above-mentioned external factors and observed associations with age at sample collection, breastfeeding, geography and antibiotics consistent with previous studies (**Supplementary Table 3**)[38,75,76]. Children living outside cities harbored more rich microbiomes compared to children in urban households throughout the first three years (q-value=0.025, **Supplementary Figure 1B**), confirming that microbial exposures from rural environments are directly reflected in the gut. We confirmed that this association was not confounded by differences between the countries (i.e., higher rate of c-sections and otherwise contrasting microbiome in Russian) by conducting the analysis without Russians and found that this association persisted within Finnish and Estonian children (linear mixed effects model, nominal p = 0.0002).

We found several associations between the microbiome and linear growth [an increase in the length (height) of a child]. Height at age three (linear mixed effects model, q-value=0.097, **Supplementary Figure 1C**) and growth rate (average increase in height per year) during the first three years of life (q-value=0.10) were associated with microbial diversity; taller and faster growing children had higher diversity trajectories throughout the three year follow-up, suggesting a link between the gut microbiome and linear growth in early childhood. On a taxonomic level, the height (q-value=0.0025, **Supplementary Figure 1D**) and weight (q-value=0.0047) at age three and height (q-value=0.012) and weight (q-value=0.00080) gain during the first three years were positively correlated with the relative abundance of genus *Dialister*. An earlier study found that malnourished Bangladeshi children (with weight-for-height Z-scores below -3) harbored immature gut microbiota[52]. Another case-control comparison between Indian children with stunted and normal growth found differences in their gut microbiota[77]. In Europe, a study found associations between the early gut microbiota at age three months and BMI at age 5-6 years in children from Finland and the Netherlands. These differences were stronger among children with a history of antibiotic use[78]. Indeed, early antibiotic use has been associated with growth in livestock, animal models and humans[79], an effect which is likely at least partially mediated through the gut microbiome[80]. The associations between early growth and the microbiome in the DIABIMMUNE cohort support the hypothesis that the early gut microbiome is an important factor in normal growth during infancy and early childhood. All in all, the findings of our association analyses, summarized in **Supplementary Tables 2, 3** and **4**, contribute to the understanding of early microbial colonization and community assembly in the human gut.

Supplementary Note 2

We isolated and sequenced eight *Bacteroides dorei* strains—three of these isolates were from two DIABIMMUNE stool samples (including two different isolates from a single stool sample) and

five from adult stool samples from the Prospective Registry in IBD Study at Massachusetts General Hospital (PRISM) cohort—using PacBio long read sequencing. This data enabled assembly of high-quality genomes, and when merged with seven existing NCBI isolate genomes, they expanded the species' pangenome (the collection of genes or gene families found in the genomes of a given species) by 7,828 genes to almost 18,000 unique gene families (**Supplementary Table 6**). 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 (**Supplementary Table 6**). For all 15 *B. dorei* isolates, this variability translated to an average of 70% inter-strain similarity in gene content, which is considerably lower than the observed SNP-based similarity (**Supplementary Figure 2B**). 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, $P<0.01$) (**Supplementary Figure 2C**). Five of these were encoded on contigs that could be circularized, suggesting an episomal entity, likely a plasmid (**Supplementary Table 7**). Anecdotally, similarity between *B. dorei* strains isolated from the DIABIMMUNE samples was greater than similarity between *B. dorei* strains isolated from adults. DIABIMMUNE *B. dorei* strains had 91% similarity between isolates from the same individual and 83-89% similarity between isolates from different children from different countries (**Supplementary Figure 2B**). In comparison, *B. dorei* isolates from adults in the PRISM cohort were on average 68% genetically similar, suggesting that the gut microbiome later in life is inhabited by more genetically diverse *B. dorei* strains and likely reflecting more heterogeneous dietary regimes and lifestyles[81]. Whether that happens through strain replacement or evolution and/or adaptation of early colonizers remains unclear.

In comparison to the pangenome constructed from the isolate genomes, 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, suggesting 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 the other species as well.

The metagenomic pangenomes generated in this study were assembled from short read Illumina sequencing data. Among the alternative approaches are long-read sequencing technologies, particularly PacBio, that produce ultra-long assemblies and high-quality draft genomes. To date this technology has been used to aid scaffolding of short-read (typically Illumina) based genome assemblies[82]. Direct applicability of the long read sequencing technologies to the traditional (bulk) metagenomics is however limited by its cost and throughput[83]. If a typical long read sequencing method (e.g., PacBio) was applied directly to a metagenomic sample, few of the most abundant species in a sample would be sequenced at full-genome breadth. Potentially, sample segregation into species-specific sub-samples using, for instance, cell sorting would allow a broader coverage of the population, though they would likely compromise on the concentration of the DNA, which is required to be high for long read sequencing technologies. In addition, current extraction methods aim to lyse a diversity of species often resulting in smaller, overly-sheared, DNA fragments from easily lysed species and longer fragments from others. Thus, obtaining representative high molecular weight DNA from a mixed community remains a large hurdle.

Supplementary Note 3

We detected 42,412 CRISPR spacers using Crass[45], a software for the identification and reconstruction of CRISPR loci from unassembled metagenomic samples, for a subset of

DIABIMMUNE stool samples (n=112) from 22 subjects with additional virome sequencing data[40]. On average, we found 382 ± 188 (SD) spacers and 29 ± 11 (SD) repeats per sample, with a positive correlation (Spearman's r=0.26; p=0.005) between number of spacers per sample and the subject age at sample collection (**Supplementary Figure 3**). This positive association disappeared after controlling for the sample alpha diversity (Shannon index), suggesting that higher CRISPR counts in older subjects result from the introduction of taxa that carry CRISPR loci rather than the acquisition of spacer sequences into the existing CRISPR loci in the infants' communities. We determined the fraction of CRISPR spacer sequences mapping to the viral contigs (indicating immunity against these viruses) assembled in a previous virome study[40]. In total, we found matching viral contigs for 2,463 (5.8%) spacer sequences, and the vast majority (n=2,085, 85%) of these viral contigs were phages (**Supplementary Figure 4, Supplementary Table 8**). Specifically, 217 spacers matched to viral contigs from the same individual, and 6 spacers matched to contigs from the same stool sample. To gain insight into the taxonomic annotation of the bacteria carrying CRISPR spacers in their CRISPR cassette, we mapped the 42,412 spacer and 3,272 repeats to the full assembly of the DIABIMMUNE dataset for the 112 samples with virome data. For 93% (n=2,285) of the spacers with a match in the virome dataset, we also found a match to the DIABIMMUNE assembly. To taxonomically annotate CRISPR-cassette carriers, we identified contigs with CRISPR cassettes including spacers matching the virome data sets by filtering all contigs without repeat matches or spacers not mapping to the virome data, resulting in 658 spacers matching contigs with one or more repeats from 32 different taxa (**Supplementary Table 9**). We observed a positive correlation between CRISPR spacer frequency and the average relative abundance of the carrier species (Spearman's R=0.452); however, some species, such as *B. vulgatus* and *F. prausnitzii*, tended to have more spacers than expected by this trend alone (**Supplementary Figure 4B**).

Members of the genus *Bacteroides* are highly versatile carbohydrate-utilizers, typically representing a large proportion of the healthy gut microbiome throughout life[84]. 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 *Bacteroides* ability to adapt their carbohydrate-active enzyme repertoire to the available resources determined by host diet. *Bacteroides*-targeting phages are among the most common members of the human gut virome, providing a plausible mechanism for extensive LGT and genomic plasticity in *Bacteroides*[85]. Indeed, phages enable LGT, for example, between *Staphylococcus aureus* strains[86] and within the Enterobacteriaceae family[87], and the study defining the virome in a subset of DIABIMMUNE samples found co-occurrence between multiple viral contigs and *Bacteroides* spp.[40]. Similarly, the most abundant members of the human gut virome, crAss-like phages, were recently associated with Bacteroidetes, especially *Bacteroides* spp.[88]. We showed that *Bacteroides* carried CRISPR spacers targeting phages in the corresponding subjects' guts identified by virome sequencing. While the CRISPR-Cas adaptive immune system has been shown to limit LGT[89-91], more recent studies have shown that CRISPR-Cas-mediated immunity enhanced LGT[92]. Here we used CRISPR spacers as a proxy to reveal phage-bacteria interactions that correlate with extensive evidence of LGT in *Bacteroides,* suggesting that phages contribute to genomic diversity in this genus.

Supplementary Note 4

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[22]. 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 (**Supplementary Figure 5A, Supplementary Table 10**). 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 (**Supplementary Figure 5B**). There was a reciprocal longitudinal dissipation of vaginal and skin bacteria (**Supplementary Figure 5C, D**), which were commonly seen in higher abundances during the first months of life.

Bacteria typical to the oral cavity are found in infant guts more often compared to adults[28]. Indeed, bacteria typical of the oral cavity in HMP spiked during the first year of life in Finnish and Estonian infants (**Supplementary Figure 5E**). Bacteria in this strata included common opportunistic pathogens (pathobionts) such as members of genera *Veillonella*, *Haemophilus* and *Streptococcus* (**Supplementary Table 10**), many of which have also been isolated from the upper gastrointestinal tracts of elderly adults[93]. We used 16S sequencing data to confirm that these genera were more abundant in Finnish and Estonian infants compared Russian infants during the same time period (**Supplementary Table 4**). 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[94]. 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[95].

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 clades[22]. 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 adult oral *V. parvula* strains found on buccal

mucosa and dental plaque but distinct from a more diverse clade typical of adult tongue microbiome (**Supplementary Figure 5F, G**). 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 (**Supplementary Figure 5H, I**). In infants, genera *Veillonella* and *Haemophilus* have been associated with formula feeding and different human milk oligosaccharide structures[96,97]. These observations demonstrate strain-level differences in oral bacteria colonizing the infant gut in relation to the adult oral microbiome.

References for Supplementary Notes

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

76 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).

77 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).

78 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).

79 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).

80 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).

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

82 Frank, J. A. *et al.* Improved metagenome assemblies and taxonomic binning using long-read circular consensus sequence data. *Sci Rep* **6**, 25373, doi:10.1038/srep25373 (2016).

83 Driscoll, C. B., Otten, T. G., Brown, N. M. & Dreher, T. W. Towards long-read metagenomics: complete assembly of three novel genomes from bacteria dependent on a diazotrophic cyanobacterium in a freshwater lake co-culture. *Stand Genomic Sci* **12**, 9, doi:10.1186/s40793-017-0224-8 (2017).

84 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).

85 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).

86 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).

87 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).

88 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).

89 Marraffini, L. A. & Sontheimer, E. J. CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science* **322**, 1843-1845, doi:10.1126/science.1165771 (2008).

90 Bikard, D., Hatoum-Aslan, A., Mucida, D. & Marraffini, L. A. CRISPR interference can prevent natural transformation and virulence acquisition during in vivo bacterial infection. *Cell Host Microbe* **12**, 177-186, doi:10.1016/j.chom.2012.06.003 (2012).

91 Palmer, K. L. & Gilmore, M. S. Multidrug-resistant enterococci lack CRISPR-cas. *MBio* **1**, doi:10.1128/mBio.00227-10 (2010).

92 Watson, B. N. J., Staals, R. H. J. & Fineran, P. C. CRISPR-Cas-Mediated Phage Resistance Enhances Horizontal Gene Transfer by Transduction. *MBio* **9**, doi:10.1128/mBio.02406-17 (2018).

93 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).

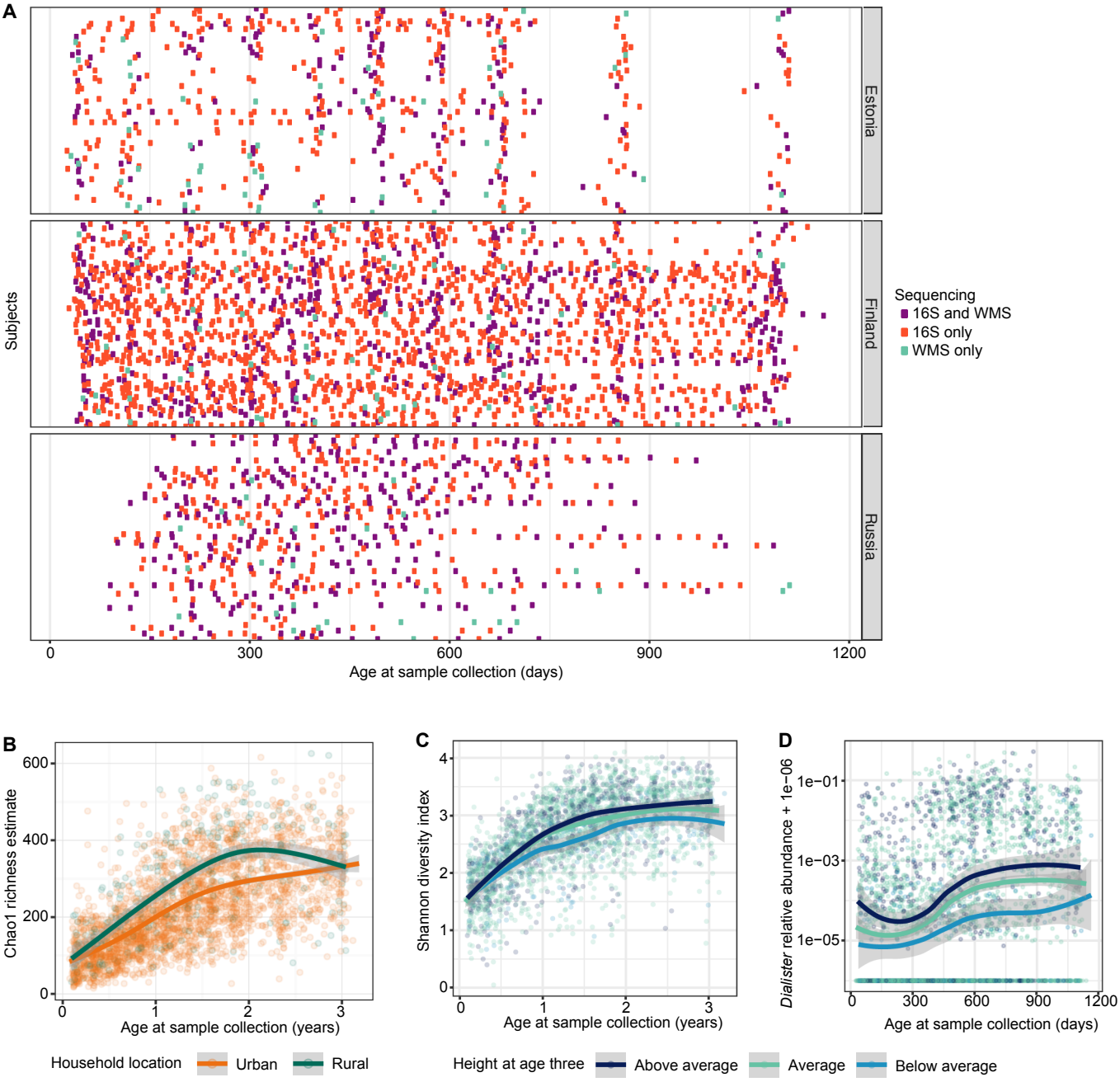
94 Buffie, C. G. & Pamer, E. G. Microbiota-mediated colonization resistance against intestinal pathogens. *Nat Rev Immunol* **13**, 790-801, doi:10.1038/nri3535 (2013).

95 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).

96 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).

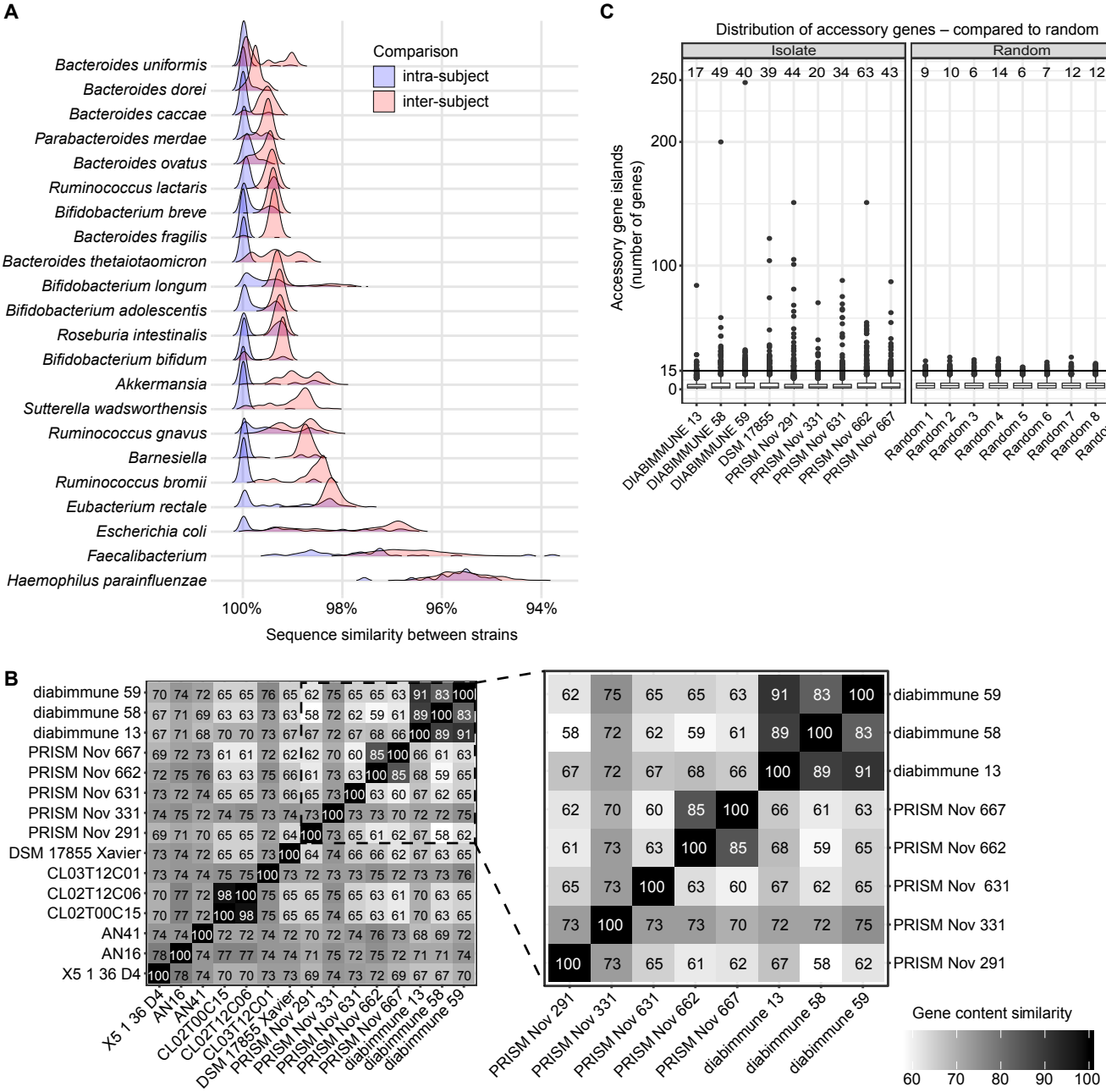
97 Guaraldi, F. & Salvatori, G. Effect of breast and formula feeding on gut microbiota shaping in newborns. *Front Cell Infect Microbiol* **2**, 94, doi:10.3389/fcimb.2012.00094 (2012).

Supplementary Figure 1



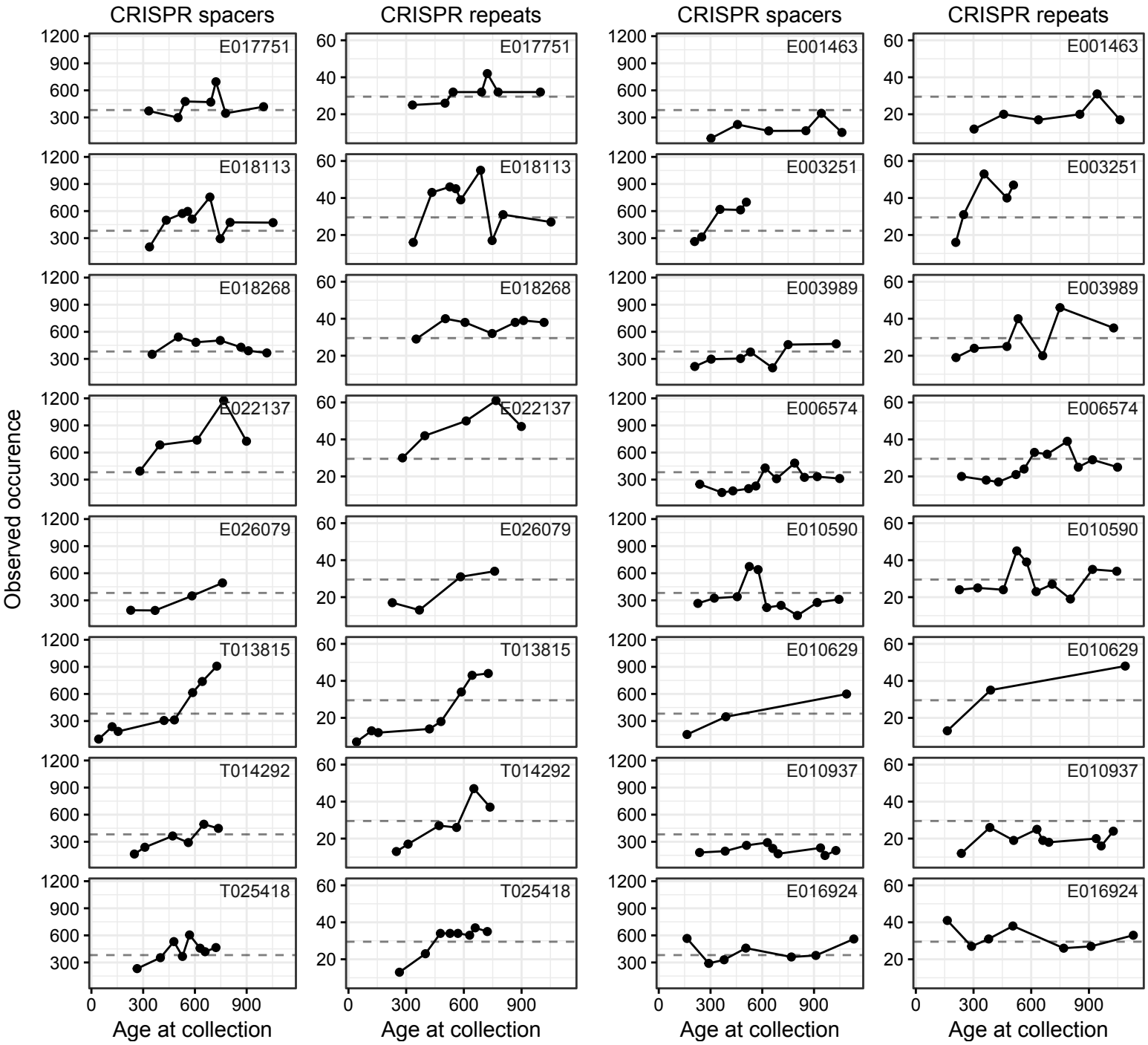
Supplementary Figure 1. A Samples analyzed by 16S rRNA amplicon and metagenomic sequencing (WMS). Rows represent subjects. **B** Chao1 richness of the microbial profiles with respect to age stratified by the household location (n=2,668 samples from urban and n=318 samples from rural households). Children born in rural households harbor more diverse gut microbiota (q-value = 0.025). **C** Children's height at the age of three is correlated with gut microbial diversity (q-value = 0.097). **D** Mean relative abundance of *Dialister* spp. in 16S sequencing profiles longitudinally stratified by subjects' height at age three. Height categories in the illustrations **C** and **D** were defined as follows. Above average: height z-score > 1, average: -1 < height z-score ≤ 1, below average: height z-score ≤ -1. In **C** and **D**, n=213 in "Below average", n=1,911 in "Average" and n=664 in "Above average" categories. In **B-D**, the curves show LOESS fit for the relative abundances and shaded area shows 95% confidence interval for each fit, as implemented in geom_smooth() function in ggplot2 R package.

Supplementary Figure 2



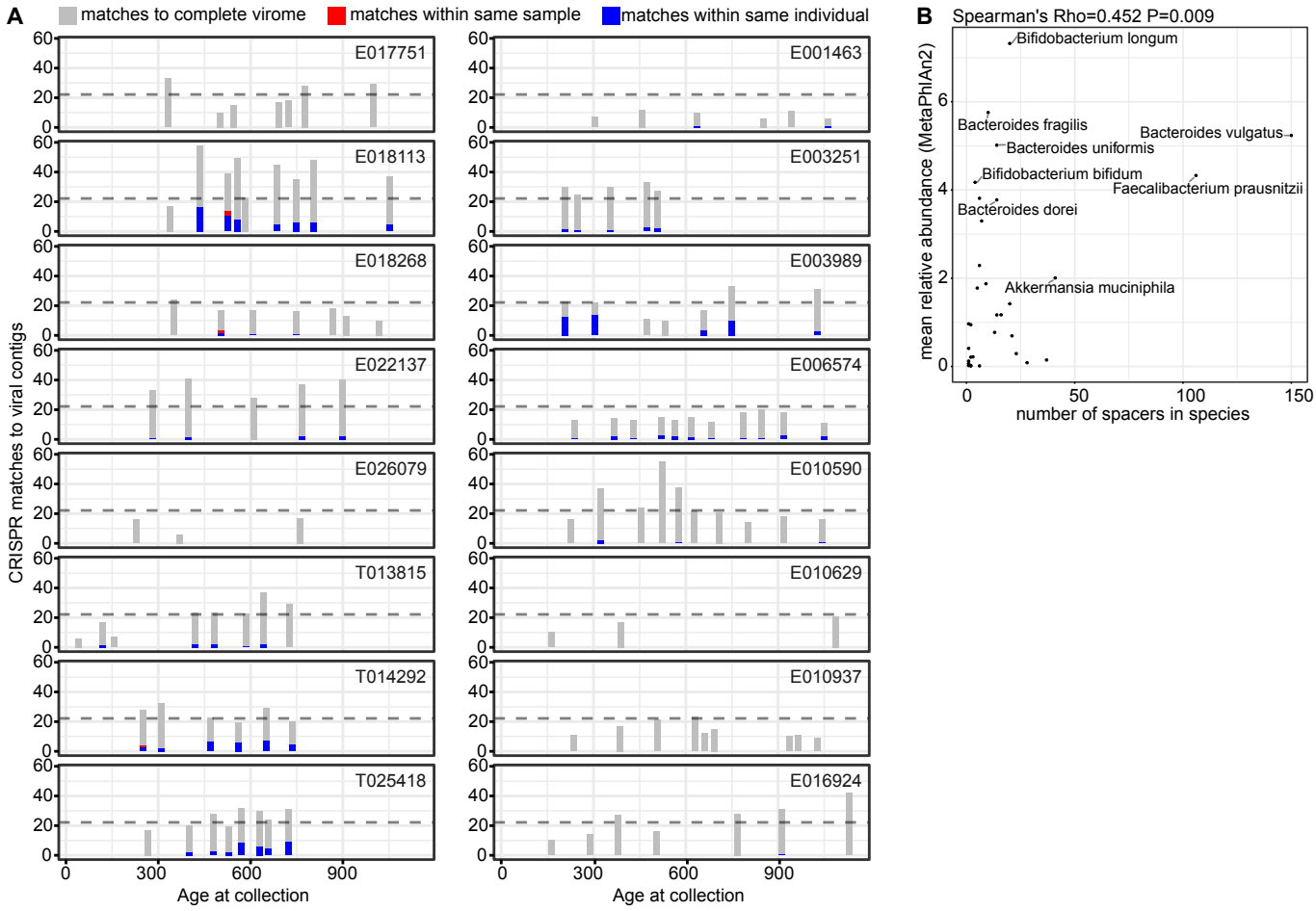
Supplementary Figure 2. Strain diversity across species and in *B. dorei.* **A** Density plots of the SNP haplotype similarities per species based on all pairwise comparisons (dominant strain per species per sample) and stratified to intra-subject (blue) and inter-subject (red) comparisons; data in Fig. 2A represented as a density plot. **B** Pairwise gene content similarities for *B. dorei* isolate genomes sequenced and assembled in this study and NCBI reference genomes available at the time of writing. **C** Distribution of accessory gene islands (adjacent accessory genes) in *B. dorei* isolates genomes from this study (left panel) is compared with randomly distributed accessory genes (right panel). Number above each boxplot indicates count of gene islands with at least 15 adjacent accessory genes (one-sided permutation test, $p < 0.05$). 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).

Supplementary Figure 3



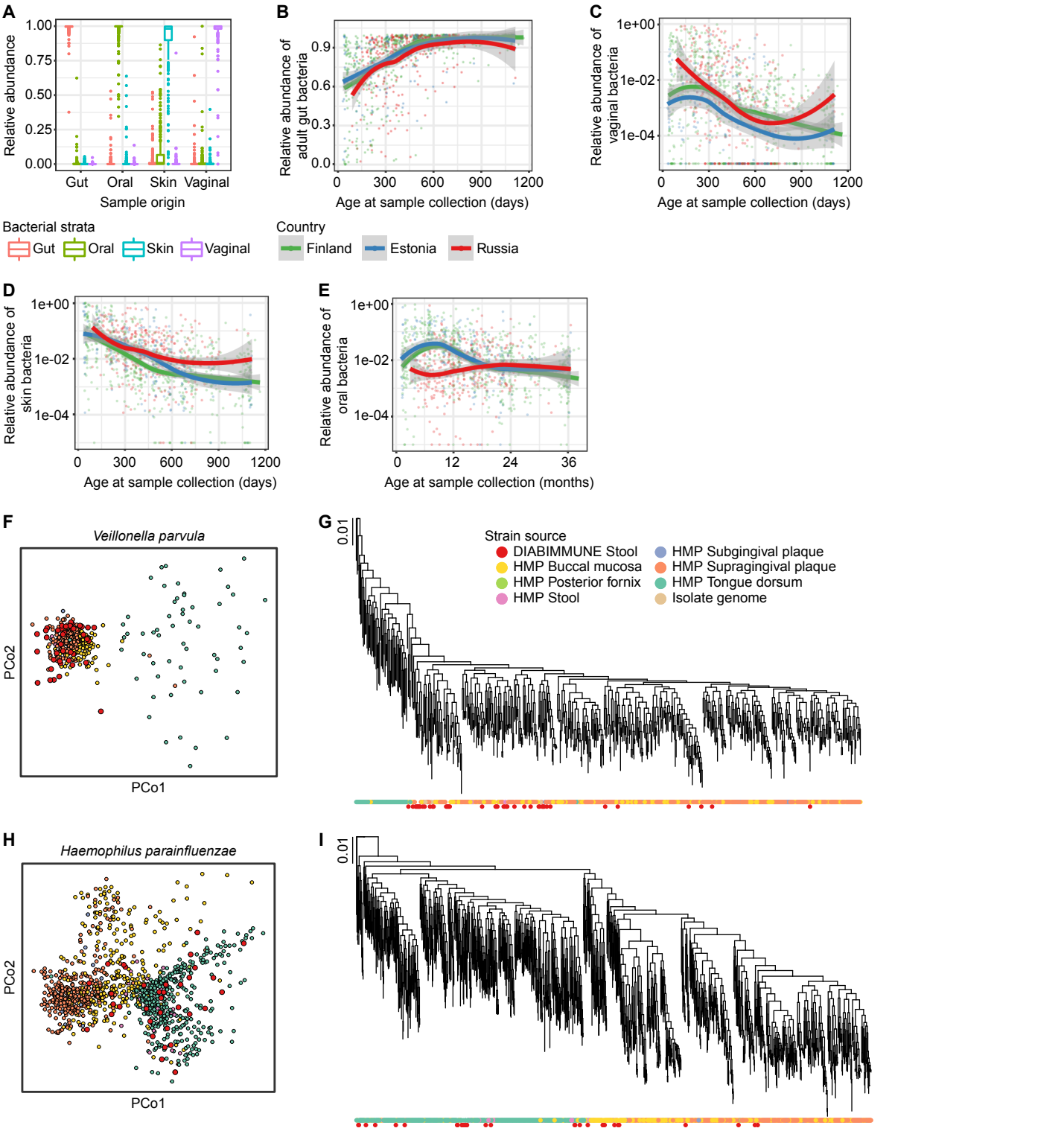
Supplementary Figure 3. Total number of CRISPR spacers correlates with age at sample collection. Individual plots display the association between age at sample collection and the number of identified CRISPR spacers and CRISPR repeats for individuals. The dashed line shows the average number of CRISPR spacers and repeats across all samples.

Supplementary Figure 4



Supplementary Figure 4. Number of CRISPR spacers matching the gut virome per individual. A Each subplot shows one individual. Grey bars indicate the number of spacers with mapping to the global virome contigs. Blue bars indicate the number of spacers (n=2,240) with mapping to the virome contigs that are drawn from samples of the same subject. Red and blue bars indicate the number of spacers that map to virome contigs that are drawn from the same sample (n=6) and from the same individual (n=217), respectively. Dashed horizontal lines show the average number of CRISPR spacer matches across all samples. **B** Positive Spearman correlation between mean relative abundance and observed number of CRISPR spacers per species (n=32 species).

Supplementary Figure 5



Supplementary Figure 5. Co-analysis of the adult microbiomes in HMP and the early 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. N=2,212 gut samples, N=5036 oral samples, N=1,236 skin samples, N=936 vaginal samples. 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). See also Supplementary Table 7. **B-E** Total relative abundance of bacterial species in longitudinal DIABIMMUNE samples classified as adult HMP (**B**) gut, (**C**) vaginal, (**D**) skin and (**E**) oral species. The color shows the country of origin and the curves show LOESS fit as detailed in the Supplementary Figure 1 legend. N=1,154 samples. **F-I** Ordination and phylogenetic tree of (**F**, **G**) *Veillonella parvula* (N=562 SNP haplotypes) and (**H**, **I**) *Haemophilus parainfluenzae* (N=1,100 SNP haplotypes) SNP haplotypes in HMP and DIABIMMUNE samples.

Supplementary Table Legends

Supplementary Table 1. Cohort metadata. This Excel file provides sample and subject metadata. Detailed description of data are given on Sheet 1 (*Table legend*).

Supplementary Table 2. PERMANOVA results. Multiple extrinsic and intrinsic factors were analyzed for connections with microbial composition using PERMANOVA. See *PERMANOVA Descriptions* sheet for details. Number of samples per test (N) is given on a separate column.

Supplementary Table 3. Microbial alpha-diversity. Multiple extrinsic and intrinsic factors were analyzed for connections with microbial alpha-diversity using mixed effects linear modeling. P-values are determined using two-sided test based on the t-statistic and corrected for multiple testing using Benjamini-Hochberg procedure. See *Alpha div. Tests Descriptions* sheet for details.

Supplementary Table 4. 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. Sample size (N) per test is shown as a separate column.

Supplementary Table 5. Strain diversity of gut microbial species. Diversity of strains within microbial species were analyzed by SNP haplotyping and gene content on metagenomic assemblies. This table supplements **Fig. 1A-C** with additional statistics. MSA = multiple sequence alignment; *MSA length* gives the effective length of the SNP haplotypes per species.

Supplementary Table 6. 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) or absence (0) of each isolate is shown.

Supplementary Table 7. Tentative circular genomic elements in the sequenced *B. dorei* isolates. Circularity was predicted by identifying highly similar regions at the start and end of a contig using Sprai (http://zombie.cb.k.u-tokyo.ac.jp/sprai/index.html). Location of overlapping regions, identity and numbers of genes are indicated.

Supplementary Table 8. CRISPR Spacer mapping to virome contigs and DIABIMMUNE assembly. Crass-derived CRISPR spacers that map to virome contigs. All spacers are shown (rows) that are found on the subset of 112 DIABIMMUNE samples that have a match to the associated virome contigs. Information of the origin of these CRISPR spacers are shown in blue, while informations to the target on the viral contigs are shown in green, and bowtie2 statistics of these alignments are displayed in grey. Mapping informations of spacers to DIABIMMUNE assembly of these samples are shown in orange which was used to infer the putative taxa of the CRISPR array carrier.

Supplementary Table 9. Most frequent taxa assigned to CRISPR spacer carrier contigs with matches to virome contigs of the DIABIMMUNE assembly. 138 spacers that match to the virome contigs were found to belong to CRISPR arrays carried from *Bacteroides* species (red).

Supplementary Table 10. Bacterial species by body site. Mean relative abundance of bacterial species in HMP data in four body sites (adult gut, skin, oral cavity, or vagina) and in DIABIMMUNE gut communities. Each DIABIMMUNE species was assigned to a body site given by the highest mean relative abundance in HMP data.

Supplementary Table 11. 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 given GO term) of 365 biological process GO terms. This table 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.