Title:

Strain-level characterization of health-associated bacterial consortia that colonize the human gut during infancy

Authors:

- Samuel S. Minot\(^1\) (sminot@fredhutch.org)
- Koshlan Mayer-Blackwell\(^2\) (kmayerbl@scharp.org)
- Andrew Fiore-Gartland\(^2\) (agartlan@fredhutch.org)
- Andrew Johnson\(^2\) (amjohns3@fredhutch.org)
- Steven Self\(^2\) (sgself@uw.edu)
- Parveen Bhatti\(^4,5,6\) (pbhatti@bccrc.ca)
- Lena Yao\(^2\) (lenayao07@gmail.com)
- Lili Liu\(^7\) (lililiu@noihp.chinacdc.cn)
- Xin Sun\(^8\) (sunxin@niohp.chinacdc.cn)
- Yi Jinf\(^9\) (13927733228@139.com)
- *James Kublin\(^2,3\) (jkublin@fredhutch.org)

* Corresponding author

Affiliations:

1. Data Core, Fred Hutchinson Cancer Center, Seattle, USA
2. Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Center, Seattle, USA
3. HIV Vaccine Trials Network, Fred Hutchinson Cancer Center, Seattle, USA

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
Abstract (250 words):

**Background:** The human gut microbiome develops rapidly during infancy, a key window of development coinciding with maturation of the adaptive immune system. However, little is known of the microbiome growth dynamics over the first few months of life and whether there are any generalizable patterns across human populations. We performed metagenomic sequencing on stool samples (n=94) from a cohort of infants (n=15) at monthly intervals in the first six months of life, augmenting our dataset with seven published studies for a total of 4,441 metagenomes from 1,162 infants.

**Results:** Strain-level *de novo* analysis was used to identify 592 of the most abundant organisms in the infant gut microbiome. Previously unrecognized consortia were identified which exhibited highly correlated abundances across samples and were composed of diverse species.
spanning multiple genera. Analysis of a cohort of infants with cystic fibrosis identified one such novel consortium of diverse \textit{Enterobacterales} which was positively correlated with weight gain.

While all studies showed an increased community stability during the first year of life, microbial dynamics varied widely in the first few months of life, both by study and by individual.

\textbf{Conclusion:} By augmenting published metagenomic datasets with data from a newly established cohort we were able to identify novel groups of organisms that are correlated with measures of robust human development. We hypothesize that the presence of these groups may impact human health in aggregate in ways that individual species may not in isolation.

\textbf{Keywords (3-10):}
Microbiome, Metagenomics, Human Development, Bacterial Consortia

\textbf{Background}

Early-life colonization of the human gut by microorganisms can have long-term implications for physiology and disease\cite{1-3}. Species- and strain-level analyses suggest that most taxa can be inherited from the mother during vaginal birth, and microbial transfer is likely reduced in infants born by Caesarean delivery or by those treated with antibiotics\cite{4-6}. Disruptions to natural bacterial exposures and microbiome development (e.g., by Caesarian section delivery, excessively sterile environment, or antibiotic-treatment) are associated with increased susceptibility to inflammatory and metabolic diseases, and intervention studies in animal models have defined key pre- and post-natal developmental windows during which the developing microbiome affects important immune processes, such as tolerance induction.
Key knowledge gaps remain concerning the immune phenotypes of at-risk infant populations and how early-life complications, such as microbiome disruption, malnutrition, and pathogen exposures, alter immune ontogeny and lead to vaccine response deficiencies in some children.

Emerging evidence suggests that individual variation in response to infection or vaccination may be influenced by past viral and bacterial exposures, which shape the immune system and can establish pre-existing immune-reactivity[7-10].

Murine systems and longitudinal human birth cohorts have defined critical neonatal windows in which the intestinal microbiome stimulates immune maturation and provides colonization resistance to protect against infectious and immune-mediated disease[3, 11-28]. While neonatal taxa-immune pathways remain to be fully elucidated, the acquisition of Clostridiales taxa in early-life is clearly vital [2, 29-44]. Clostridiales provide colonization resistance[13], stimulate immune-regulatory responses[18, 26, 45-47], and activate IFN-mediated lung protection[48]. A failure to acquire Clostridiales taxa, especially Ruminococcaceae, Lachnospiraceae and Clostridium Cluster XIVa, represents the major deficiency of the CF infant microbiome, a finding that is highly reproducible across multiple independent cohorts, including the most extensively characterized BONUS cohort[30, 31, 35, 37, 44].

Longitudinal studies of birth cohorts – so far conducted predominantly in North America and Europe – have begun to characterize compositional changes to the gut microbiome that occur in the first years of life. These studies have relied primarily on amplification of the bacterial 16S
ribosomal RNA gene or, more seldomly, whole genome sequencing[4, 6, 49-51]. These longitudinal studies, along with one major cross-sectional study[52], have demonstrated that there is considerable inter-individual and temporal variation in the neonatal and infant microbiome community starting from birth and extending to approximately three years of age. During this period the microbiome gains richness and stability to form a microbial community that is more reflective of the adult microbiome[4, 6, 49-51, 53]. This represents a general transition where bacteria specialized to the aerobic neonatal gut (e.g., *E. coli*) or for growth on complex sugars in breastmilk (e.g., Bifidobacterium and Veillonella) and are outcompeted by organisms found more commonly in the adult gut microbiome, such as Bacteroideceae and Rumminococcaceae[4, 6, 50]. This is reflected in the metagenomic composition of the bacterial communities, with genes involved in milk oligosaccharide metabolism giving way to those better suited to solid foods, such as fiber degradation[4, 6, 49].

While these studies have elucidated general trends in infant microbiome development, most prior studies are limited by low density of fecal sampling in the first 6 months of life when temporal intraindividual variation in the microbiome is highest and exposures to the immune system are particularly impactful. Furthermore, bioinformatic approaches have focused predominantly on identifying microbiome community states that are reflective of specific ages and which are generalizable across individuals. In contrast, these approaches offer more limited insight into the growth dynamics of individual taxa or clusters of interacting consortia. It is now evident that the path of individual microbiome development is highly variable across infants. For example, a recent analysis of transitions between ten different microbiome community
states in early life observed great diversity in the patterns of transitions between states. In fact, the most common transition pattern was only observed in 20% percent of infants, with the remaining 80% of infants displaying unique maturation transition patterns.[54]

Recognizing that the microbiome may not conform to consistent community states, an alternate ontological approach is to identify the subsets of microbial organisms which are found together in concert, as would be expected from a group of Bifidobacteria which jointly metabolize human milk oligosaccharides. From an informatic point of view, we approach the identification of microbial consortia by testing for organisms with correlated abundances across large numbers of microbiome samples[55]. When organisms are more likely to be found together than would be expected by random assortment, we may hypothesize that there is a shared underlying biological process which is jointly driving their growth and survival.

In human microbiome research, detection of well-studied bacterial species and genera can reveal considerable information about environmental conditions present during health and disease conditions of the host; however, there are limitations in taxonomic-centered approaches that orient analysis around finding associations with host characteristics and relative abundance of bacterial groups agglomerated by phylogenetic clade. Critically, aggregating organismal relative abundances within phylogenetic clades (i.e., summarizing microbiome features to the genus, family, or order level) becomes less informative as physiology and metabolism of bacteria within taxonomically derived grouping can vary greatly. However, strain-level analysis suffers from high-dimensionality and high sparsity of features.
between samples. Thus, a major challenge in analysis of microbiome is finding a flexible unit of analysis that permits detection of consistent and interpretable ecological changes in the host via phylogenetic-independent agglomeration of co-abundant organisms.

Thus, to gain a better understanding of dynamics of microbiome development in the critical development period between 1 to 6 months of age, we conducted such a gene-level microbiome analyses on stool samples collected monthly in a longitudinal mother-infant birth cohort\[56\]. Because the aggregate gene content of the gut microbiome is comprised of tens or hundreds of millions of genes\[57\], a meaningful embedding in lower dimensional space is helpful for comparisons across samples. For this purpose, we use Co-Abundant Gene Groups (CAGs)\[58, 59\] which represent sets of genes that are expected to be found together in the same genetic element (chromosome, plasmid, virus, etc.) across all of the samples in the collection. To increase the total amount of biological information used for CAG construction, we augmented the data from our own cohort with additional metagenomic data from seven published infant microbiome datasets\[4, 6, 49-53\], for a total of 4,441 biological samples in the combined dataset. We used a reproducible pipeline, geneshot [60], for constructing and quantifying CAGs to describe groups of organisms that colonize the gut according to patterns of correlated abundance which are reproduced across cohorts. We describe these groups of microbes which are present or absent in concert as previously-unrecognized microbial consortia, which may help researchers more succinctly describe the patterns of rapid turnover which are observed during the first six months of life. Moreover, parallel analysis of a published cross-sectional cohort\[53\] identified one such consortium whose presence was strongly
correlated with infant growth rate. We propose that this dynamic growth variation may
underlie altered immune development between individuals and associated susceptibility to
immune-mediated disease in later life, and therefore that CAG-based analysis of microbial
consortia would be a useful approach for the analysis of existing and future longitudinal birth
cohorts.

Results

De novo metagenomic analysis identifies 592 bacterial genomospecies in the infant gut
To quantify the relative abundance of microorganisms present in the gut during early human
life, metagenomic whole-genome shotgun sequencing (WGS) data were generated from the
total DNA isolated from stool samples collected from cohort of healthy infants (n = 15) at
monthly intervals from birth until 6 months. To mitigate the inherent stochasticity of
metagenomic sequencing, we sought to increase generalizability of gene-level analysis by
including other publicly available deeply sequenced microbiome samples from other infant
cohorts. In all, we analyzed metagenomic data from the seven largest
published infant microbiome datasets[4, 6, 49-53], for a total of 4,441 biological samples in the
combined dataset (Fig. 1A, Table 1, Table S1, S2). Gene-level metagenomic analysis was
performed by: (1) generating a gene catalog via de novo assembly and centroid clustering
(based on 90% amino acid identity); (2) estimating the relative abundance of organisms
encoding each individual gene via short-read alignment; and (3) grouping genes with correlated
abundances into co-abundant gene groups (CAGs) via iterative greedy single linkage
clustering[60]. Thus, the primary unit of measurement for downstream analyses was the relative abundance of each CAG in a particular sample, which is an estimate of the relative abundance of the organisms contained within that sample encoding the genes in that CAG. Because the groups of genes contained in each CAG have highly correlated abundances, they are predicted to be contained within the same genomic context with the metagenome, representing the complete or partial core genome of a set of closely related isolates or strains[58].

To focus our analysis on CAGs most likely to represent species-level groupings of genes, we subset our analysis to those CAGs containing between 1,000 and 10,000 genes (n = 592), a range which encompasses most representative bacterial genomes in the NCBI RefSeq database (Fig. 1B). We conceptualized these CAGs as “genomospecies” because they define a group of organisms at the species or strain level based on a high degree of shared genomic content[61]. The filtered set of 592 appropriately sized CAGs, i.e., genomospecies, account for over half of the raw sequence fragments recovered from infant stool samples with no clear bias by study or timepoint (Fig. 1C). While the organisms contributing the remaining sequence fragments may also have a meaningful influence on human health, they were not observed consistently at high enough abundance across multiple samples to enable gene-level analysis in this study. Using these genomospecies-level abundances as the basis of characterizing microbiome composition in our to -cohort, we compared pairs of samples from the same or different individuals using rank correlations across the CAGs. Sample pairs from the same individual were more correlated (Fig. 1D, p=1.16E-46 Mann-Whitney U), as were samples collected from the same individual at
shorter time intervals compared to longer intervals (Fig. 1E, \( p=4.12\times10^{-10} \) Spearman); together these show that there is some degree of temporal stability in community composition. A graphical summary of microbial abundances across samples is shown in Figure 1F, with each genomospecies shown in a row and each sample shown in a column. Samples are ordered by participant and timepoint, and the genomospecies are grouped by linkage clustering based on the similarity of abundance patterns across samples. The presence of genomospecies with very similar patterns of abundance in this dataset suggests that organisms are not distributed randomly across individuals, but that there may be groups of genomospecies whose relative abundance are correlated when comparing across specimens (Fig. 1F).
Figure 1. Quantification of 592 microbial genomospecies from 0-6 months across 8 studies.

A) Density of sample collection per study over time relative to term birth (40 weeks of gestational age).

B) Distribution of genomospecies genome sizes (number of coding sequences, top) in comparison to NCBI prokaryotic reference genomes (bottom).

C) Proportion of metagenomic sequence data from each sample which can be unambiguously assigned to any one of the 592 bacterial genomospecies, compared across studies.
D) Within- versus intra-participant variation in metagenome similarity estimated using the Spearman correlation coefficient of Centered-Log Ratio (CLR) abundances.

E) Comparison of within-participant variation in metagenome similarity between samples at varying time intervals estimated using the Spearman correlation coefficient of CLR abundances.

F) Comparison of microbiome composition within the samples collected for this study, with participants (n=15) and timepoints (n=6) indicated on the top marginal axis.

Dendrogram indicates hierarchical clustering of the 592 genomospecies based on similarity of abundance profiles across samples. Color scale indicates log-scaled relative abundances.

Bacterial strains are observed in tightly correlated consortia across populations.

To identify microbial species with correlated abundances, we calculated the Kendall rank correlation coefficient[62] for every pair of genomospecies across all of the samples from both published and newly-generated microbiome samples. Ordination of genomospecies based on similarity of abundance profiles across samples suggested a correlation within taxonomic groups (Fig. 2A; Supp. Data 1). While correlation coefficients were higher overall within taxonomic groups than between taxonomic groups (ANOSIM R=0.43 p=0.001) this was not observed universally across taxonomic groups, with many CAG pairs from the same taxa showing a complete lack of correlation (Fig. 2B, S1). Having observed that taxonomic similarity was not the primary driver of correlated abundances, we compared all pairs of genomospecies in a taxonomically-agnostic analysis. Out of all pairwise comparisons of genomospecies...
(n=174,936) only 143 had a Kendall’s tau value of 0.7 or greater (shown with connecting lines in Fig. 2A). The correlation of this subset of CAG pairs remained strong even after filtering the data to only include a single sample from each participant or calculating the correlation independently for each study (Fig. 2C), and so it is not likely to be driven by the confounding effect of intra-individual or inter-population differences in community composition.

Next, groups of microbes were identified by single-linkage clustering using these highly correlated genomospecies, which we conceptualized as “consortia” because of their high degree of co-abundance in the infant gut microbiome (Table S3). We retained all genomospecies in this analysis, with those that did not have any correlated match being treated as individual single-genomospecies “consortia.” The most abundant consortia accounted for 1-5% of all predicted genome copies on average across all specimens in the meta-analysis (Table S4, S5). To test our hypothesis that these highly correlated genomospecies represent multiple organisms (in comparison to the null hypothesis that that a single organism encodes all of the observed co-abundant genes), we compared these metagenome-derived genomospecies to the reference genomes of bacterial isolates. To identify the bacterial reference genomes which are most similar to each genomospecies we searched the NCBI RefSeq collection of bacterial genomes (n=113,938; downloaded June 6th, 2022) by amino acid sequence alignment. To better understand genomospecies relationships to conventional phylogenetic based metagenome interpretation, we closely examined the two largest groups of genomospecies with highly correlated abundances (Fig. 2D, S2). We made two observations. First, the genes contained within each individual genomospecies generally mapped to a
consistent set of genomes. Second, genomospecies within those consortia often mapped to
different strains and species within a genus (Fig. 2A,2E) or even different orders within a class
(Fig. 2F, Supp. Data 2). Finding no single genome with the complete genetic content present in
these correlated genomospecies, it is likely that they represent groups of distinct organisms
that are present at correlated relative abundances in the human gut microbiome during early
life.
Figure 2. Groups of microbial genospecies are reproducibly observed at correlated abundances across studies.
A) UMAP-based ordination of genomospecies (filtered at a minimum threshold of 0.1% average abundance) based on correlation of relative abundances across samples (Kendall’s tau).

B) Comparison of relative abundance-based correlation coefficients for genomospecies pairs based on order-level taxonomic annotations.

C) Considering only those pairs of genomospecies with a correlation coefficient greater than 0.7 (Kendall’s tau) using all available data, correlation metrics were recalculated using a single timepoint per participant across all studies (“Single Timepoint”); while calculating an independent correlation metric for each individual study (“Within Study”); or using a single timepoint per participant while also calculating an independent correlation metric for each individual study (“Within Study – Single Timepoint”). The Spearman correlation coefficient was also calculated for all comparisons (blue) in addition to Kendall’s tau (orange).

D) Comparison of relative abundances (CLR) across all samples for each pair of genomospecies within Consortium 5.

E) Bacterial reference genome similarity for each of the genes within the 4 genomospecies which make up Consortium 5. Each column represents a single gene reconstructed from the metagenomic analysis. The bottom color bar indicates the genomospecies (CAG) assignment for each gene. Blue marks indicate reference genomes (each shown in a distinct row) in which that gene was detected by sequence alignment. The right-hand color bar indicates the species-level assignment for each reference genome. Hierarchical
clustering of reference genomes is based on the average nucleotide identity-based
dissimilarity matrix.

F) Bacterial reference genome similarity for Consortium 3 (following E), with the full set of
reference genomes available for inspection in Supplementary Data 2.

Relative abundance of microbial consortia changes rapidly during human infancy

Considering the human gut microbiome as a collection of microbial consortia, we wanted to
better understand how this complex community evolves during early life. Individual consortia
vary widely in relative abundance both as a function of host age as well as study population
(Fig. 3A-B, Supp. Data 3). Because each study included samples from a single population, it was
not possible to distinguish between study population differences and batch effects of sampling.

Ordinating samples based on consortium abundances across all studies shows a complex
pattern, suggesting that samples at earlier timepoints are more varied in community
composition, and samples at later timepoints converge on a smaller number of community
types (Fig. 3C, Supp. Data 4). Consistent with this hypothesis, we found that the composition of
microbial consortia was more similar at premature-birth and later timepoints within each study
(Fig. 3D, p=0.008 Spearman).

To better understand the dynamics of microbial communities as a function of host age, we used
ANOSIM to compare all the samples collected at each pair of timepoints within each study. In
our data, we noted a similarity of days 30-60 as well as days 90-150, with day 180 as the most
distinct (Fig. 3E). In contrast, the Yassour et al. dataset showed a greater similarity of days 60-90
than 30-60 (Fig. 3F). Looking entirely at the similarity of samples from the same individual over
time, our data showed a greater degree of change (lower correlation) from days 60-90 than 30-
60 or 90-120 (Fig. 3G), while the Eng et al. data showed a greater degree of change from 30-60
than 60-90 or 90-120 (Fig. 3H). The combined analysis across cohorts emphasizes the high
degree of interpersonal heterogeneity and temporal transience in the human gut microbiome
during early life.
Figure 3. Rapid changes in relative abundance of microbial consortia during early human life.

A) Summary of the relative abundance of Consortium 5 (vertical axis) across stool samples as a function of time since term birth (horizontal axis), comparing samples obtained from different studies (indicated by color).

B) Summary of the relative abundance for Consortium 3, as in (A).
C) UMAP-based ordination of microbiome samples based on similarity of microbiome composition as measured by the relative abundance of microbial consortia. Colors indicate the timepoint of sample collection relative to term birth.

D) Similarity of sample composition was compared for pairs of samples collected at similar timepoints from different individuals within each study using Kendall’s tau. The horizontal axis indicates the time of sampling, and the vertical axis indicates the similarity of microbial abundances observed between different individuals.

E) Similarity of microbial relative abundances were compared between pairs of samples collected at different timepoints from different individuals within the samples collected for this study (Kublin). The pairwise comparison of each timepoint using the ANOSIM R metric is shown in a heatmap, with positive values indicating more distinct microbial compositions within each of the pair of timepoints and negative values indicating more similar microbial compositions within the pair of timepoints.

F) Similarity of microbial relative abundances for the samples from the Yassour study (as in E).

G) Similarity of samples collected from the same individual at adjacent timepoints within the samples collected for this study (Kublin). The horizontal axis indicates the timepoint which was compared to samples from the immediately proceeding timepoint. Higher values on the vertical axis indicate a greater similarity of samples based on the relative abundance of microbial consortia.

H) Similarity of samples collected from the same individual from the Eng study (as in G).
Taxonomically diverse microbial consortia are less abundant in the gut of infants with cystic fibrosis. An important application of detailed microbiome analysis is to identify microbes which may influence human health and disease. A study published by Eng, et al.[53] paired metagenomic sequencing with infant health indicators and compared the metabolic pathways encoded by the microbiome with inflammation and nutritional failure in We accessed a rich dataset that paired data from infants with cystic fibrosis (CF; n=207) to healthy controls (n=25). As previously observed[63], infants with CF had lower weight than healthy controls at each timepoint (Fig. 4A, Wilcoxon p=0.0017). To identify organisms with relative abundances that are correlated with CF status and/or weight, we performed independent linear modeling at each timepoint. The weight-association analysis was performed using only samples from participants with CF. Because of uneven sampling between groups, the CF- and weight-association analyses were performed over an overlapping but distinct set of days. When comparing the strength of association with these two clinical features across the consortia, we observed that the organisms with positive weight-associations (observed at higher abundances in CF infants with greater weights) generally had negative CF-associations (observed at lower abundances in CF infants compared to healthy controls), and vice versa (Fig. 4B-C).

The microbial consortium showing the strongest association with weight in the CF infants was #9 (Fig. 4D), which was also found at lower abundance in CF infants compared to healthy controls (Fig. 4E). Alignment of the genomic markers of consortium #9 against the NCBI RefSeq catalog of microbial genomes identified species spanning Bacteroides (B. uniformis, B. stercoris, B. eggerthii, B. ovatus, B. caccae, and B. thetaiotaomicron), Paraprevotella clara/xylaniphila,
and *Phocaicola* (*P. plebius* and *P. coprocola*) (Fig 4F). While these genera have been identified previously as being altered in the gut microbiome of infants with CF, these results: (a) identify the species that are most likely involved with a specific set of genomic markers (Supp. Data 5), (ii) indicate that those species are generally found together rather than individually in the gut microbiome, and (iii) suggest that the combined metabolism of a multi-species consortium may collectively mediate weight gain in infants with CF.
Figure 4. Association of specific microbial consortia with infant CF status and weight.

A) Measured weight of each infant at each timepoint, distinguishing infants diagnosed with CF from healthy controls.

B) Estimated coefficient of association for the relative abundance of each microbial consortium with CF status, calculated independently at each timepoint.
C) Estimated coefficient of association for the relative abundance of each microbial consortium with infant weight using only those participants diagnosed with CF, calculated independently at each timepoint.

D) Comparison of the relative abundance of Consortium 9 with weight within the group of participants diagnosed with CF, shown independently at each timepoint.

E) Comparison of the relative abundance of Consortium 9 between participants distinguished by CF diagnosis, shown independently at each timepoint.

F) Comparison of the genomic content of Consortium 9 to a reference genome collection, as in Figure 2E.

Discussion

Quantification of microbes sampled from the human gut

By using the reference-free analysis approach to microbiome analysis implemented in the geneshot analysis pipeline[60], our analysis aimed to expand our understanding of the infant gut microbiome. The advantage of this approach, which quantifies organisms on the basis of the genes encoded in their genome, is that it is not dependent on the composition of existing genome databases to detect and quantify specific organisms. While the primary drawback of this approach is a lack of sensitivity for the detection of organisms that are not sequenced to a depth sufficient for de novo reconstruction, approximately 50% of the raw metagenomic data was successfully assigned to just 592 distinct genomospecies representing the most abundant organisms (Fig. 1C). Based on previous work, we expected that microbial composition would reflect some degree of individuality and temporal stability[55, 64, 65]. This expectation was
borne out using the genomospecies-level abundance data, with samples more similar within-
than between-participants (Fig. 1D) and more similar between samples collected across shorter
time intervals (Fig. 1E). Based on these high-level metrics, we gained confidence that our
genomospecies-level analysis is capturing a biologically meaningful profile of the most
abundant organisms in the infant gut microbiome.

Observation of taxonomically distinct microbial consortia

In addition to the detection of previously unsequenced organisms, an advantage of de novo
metagenomic analysis is the ability to precisely identify organisms with correlated abundances
that are taxonomically similar or diverse. While marker-gene or k-mer based analyses run the
risk of confounding taxa that share a subset of genomic content, our de novo gene-level
analysis assigns each raw sequence read unambiguously to a single genomospecies reference
(using an expectation maximization approach to resolve duplicate alignments). Moreover, by
limiting to the 592 organisms, evaluating all possible pairwise correlations among microbes
became computationally tractable. Using this approach, we found only 143 pairs of microbes
(out of the 174,936 total pairwise comparisons) with a Kendall’s tau correlation coefficient ≥0.7.
Noting the inter-participant individuality of microbiome composition (Fig. 1D), we were
encouraged that this high degree of pairwise correlation between individual genomospecies
was observed after downsampling to a single sample per participant and after controlling for
batch effects (Fig. 1C).

While it is possible that genomospecies with correlated abundances may represent a single
species which was inappropriately split due to noise in the metagenomic sequencing process, it
is more likely that correlated genomospecies represent different species with correlated
abundances. One biological concept used to describe such multi-species groups would be
“consortia” of distinct organisms formed by cross-feeding or syntrophic interaction[66] or by
stable niche partitioning of a common source of energy (such as the degradation of diverse
human milk oligosaccharides by related Bifidobacteria[67]). By comparing each genomospecies’
genetic content to the extensive NCBI RefSeq genome collection, we observed candidate
consortia containing genetically distinct organisms from the same genus (e.g., Consortium 5,
Fig. 2E, Supp. Data 2), as well as single consortia containing organisms spanning multiple
diverse genera (Klebsiella, Enterobacter, Leclercia, Citrobacter, Cronobacter, Proteus, Serratia,
and Pseudomonas) (e.g, Consortium 3, Fig. 2F, Supp. Data 2). The robust correlation of relative
abundances between these genetically distinct organisms is highly unlikely to be caused by
technical artifacts, and strongly suggests that these groups of organisms are present or absent
in the microbiome as a correlated group.

Complex, rapid temporal dynamics of the infant gut microbiome

Using the aggregate abundances of microbial consortia to measure the composition of the gut
microbiome, we sought to better understand the complex temporal dynamics of the developing
human microbiome during infancy. While there were some consistent patterns across datasets
– consortium #5 of Bifidobacteria was observed at higher abundance during later timepoints
(Fig. 3A) and consortium #3 of diverse Enterobacterales observed at higher abundance during
earlier timepoints (Fig. 3B) – those patterns were not consistent across all individuals or all
studies. Clustering of samples by total community composition did not reveal any single
community state associated with earlier or later timepoints (Fig. 3C). The most consistent pattern we observed was that microbial communities from later timepoints were more similar across individuals than the communities from later timepoints, an effect which was observed across multiple independent studies (Fig. 3D) and which is consistent with the previous observations made within single cohorts [11, 54, 55].

The development of the human microbiome during the earliest days of life is a highly dynamic process which has not been measured at consistent, dense intervals across previous studies. We augmented the published set of microbiome studies by collecting stool samples at 30-day intervals from birth to day 180 in a cohort of 15 infants. While the stool microbiome in this cohort was more similar between days 30-60 and 90-150 (Fig. 3E), a previous study has shown a stronger signal of similarity between days 60-90 (Fig. 3F). To identify the patterns of microbiome development that are consistent across populations, the field will need to collect considerably more metagenomic data at higher temporal frequency from this early time period across multiple geographically diverse study sites.

Identification of a multispecies microbial consortia associated with health outcomes in infants with cystic fibrosis

To assess whether any of the newly-identified microbial consortia were correlated with human health outcomes, we focused on a study of infants with cystic fibrosis [53], performing linear modeling of consortium abundances with both CF status and weight among infants with CF (those analyses being performed independently). The strongest association was observed with a multispecies consortium (#9) containing species of *Bacteroides, Paraprevotella, and*
*Phocaeicola* (Fig. 4F). This group of organisms was found at lower abundance in infants with CF compared to healthy controls; and critically was also found at lower abundance in those infants with CF who weighed less; in particular at day 90 of life (Fig. 4B-E). Our biological interpretation, which is heavily influenced by the identification of this microbial consortium, is that there is a mechanistic link between the presence of this group of microbes with the factors influencing weight gain during early life. Similar to the joint metabolism of human milk oligosaccharides distributed across related *Bifidobacteria*[67], we hypothesize that there is a consequence from the presence of this group of organisms which may not be recapitulated by any single member. When translating these findings to a controlled experimental setting, our results would imply that the administration of any single species may not be sufficient to reproduce the same biological effect, but instead the full or partial set of the multi-species community may be required.

**Methods**

**Study sites and enrollment of cohort.**

We obtained data from a collaborative mother-infant cohort that enrolled pregnant women in the Guangdong and Zhejiang Provinces of China from and followed their newborn off-spring from birth up to two years of age[56, 68]. Samples were collected for this analysis from 12/19/2017 to 08/21/2018. Pregnant women provided written informed consent and were screened and enrolled between 14 and 20 weeks of gestation. The study completed enrollment into the cohort in January, 2020.
Specimen and data collection and processing.

Stool samples were initially collected at the hospital (10 grams of fecal matter collected from diapers, placed in plastic containers and stored at -80°C). Subsequent stool samples were collected monthly by parents in the home. On the morning of sample collection, a cooler with ice packs and sample collection materials was delivered to the home of each participant. In the early evening, the coolers were collected and returned to the laboratory where samples were aliquoted, labelled and stored at -80°C. If an infant did not provide a stool sample on the collection day, collection was rescheduled for the following day and a new cooler was provided.

Sequencing

DNA was extracted from each stool sample (n=94) and prepared for sequencing using the TruPrep DNA Library Prep Kit V2 for Illumina. Libraries were clustered and sequenced on an Illumina HiSeq2000 instrument and sequenced to an average depth of 61.7 million paired-end reads per sample.

Sequencing data from published datasets were obtained using the SRA Toolkit, downloading all paired-end FASTQ data available from the BioProject accessions PRJNA521878 (Bender), PRJEB6456 (Backhed), PRJNA630999 (Bajorek), PRJNA475246 (Yassour), PRJNA698986 (Lou), PRJNA486782 (Leonard), and PRJNA510445 (Eng).

Analysis

Identifying and quantifying CAGs from metagenomes
While previous reports have described bacteria in broad taxonomic groups (e.g. Bifidobacteriaceae, Lactobacillus, Enterococcus, Bacteroides, Streptococcus), we used gene-level metagenomics to increase this level of resolution to the species- and strain-level, while also identifying horizontally transferred genetic elements which play a role in microbiome development. Analysis of raw FASTQ datasets was performed using the geneshot analysis pipeline, available at https://github.com/Golob-Minot/genshot. The exact version of that software used was v0.9 with the commit hash 4d700993660ed8fdf4df6432d2c7cb2ddd8ce85f. The geneshot pipeline (described previously [60]) performs the following bioinformatics analysis steps:

1. De novo assembly of each sample independently (using megahit v1.2.9);
2. Identification of protein-coding sequences in each assembly (using prodigal v2.6.3);
3. Deduplication of protein-coding sequences at 90% sequencing identity and 50% coverage (using linclust/MMseqs2 release 12-113e3);
4. Alignment of conceptually translated sequence reads against that deduplicated gene catalog (using DIAMOND v0.9.10);
5. Clustering of protein-coding sequences into CAGs using a maximum cosine distance threshold of 0.35.

To effectively process the large number of metagenomic samples in this project, a subset was selected for de novo assembly and gene identification which included only a single representative per participant across all projects, while genes and CAGs were quantified across the full set of samples. The computational resources required for this analysis were considerable, with ~104,000 CPU hours required for the de novo assembly and gene identification.
identification and ~364,000 CPU hours required to align the full dataset against that gene
catalog.

Comparing the composition of microbial communities
The similarity of organisms present in different microbial communities was estimated using the
non-parametric Spearman correlation of CLR-transformed abundances. The Spearman R value
was used when comparing pairs of samples in terms of their microbial composition. When
comparing pairs of CAGs to find organisms with correlated abundances, the more conservative
Kendall Tau metric was also calculated using the CLR-transformed abundances.

Identifying genomospecies associated with health status
Statistical analysis for the association of genomospecies relative abundance with the health
status of human hosts (CF status and weight) was performed using Generalized Estimating
Equations as implemented in the statsmodels package (Python). All GEE models were
constructed using an exchangeable covariance structure and Gaussian family. Adjustment for
multiple hypothesis testing was performing with the FDR-BH protocol as implemented in
statsmodels.

Comparing genetic content of genomospecies to reference genomes
The reference genomes most closely resembling the organisms reconstructed de novo from this
metagenomic dataset were identified by alignment against the NCBI RefSeq database
(downloaded June 6, 2022). The protein-coding sequences from each CAG (“genomospecies”)
was aligned against that genome collection using the gig-map workflow (available at [https://github.com/FredHutch/gig-map/](https://github.com/FredHutch/gig-map/)), which employs the DIAMOND aligner for rapid alignment of conceptually-translated genomes in amino acid space, at a minimum alignment threshold of 90% sequence identity and 90% alignment coverage (of the *de novo* assembled gene sequence). The similarity of reference genomes (used for the dendrogram display in CAG-genome heatmaps) was estimated by gig-map with Average Nucleotide Identity (ANI) calculated using the MASH software [69].

### Tables

**Table 1**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Citation</th>
<th>Samples (#)</th>
<th>Participants (#)</th>
<th>Mean Reads per Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kublin</td>
<td>This study</td>
<td>94</td>
<td>15</td>
<td>61,722,891.79</td>
</tr>
<tr>
<td>Bender-PRJNA521878</td>
<td>1</td>
<td>62</td>
<td>29</td>
<td>4,166,371.97</td>
</tr>
<tr>
<td>Backhed-PRJEB6456</td>
<td>2</td>
<td>400</td>
<td>100</td>
<td>39,764,708.24</td>
</tr>
<tr>
<td>Nguyen-PRJNA630999</td>
<td>3</td>
<td>292</td>
<td>77</td>
<td>64,487,422.86</td>
</tr>
</tbody>
</table>
Yassour-PRJNA475246 | 4 | 169 | 43 | 28,045,378.54

Lou-PRJNA698986 | 5 | 2,049 | 642 | 26,306,197.79

Leonard-PRJNA486782 | 6 | 96 | 24 | 43,482,751.56

Eng-PRJNA510445 | 7 | 1,279 | 232 | 27,790,671.30

Summary of metagenomic WGS gut microbiome datasets included in meta-analysis.

**Supplementary Tables**

- Supplementary Table 1: Manifest with metadata for all specimens, including number of reads, number of genes detected, etc.
- Supplementary Table 2: Relative abundance of all genomes across all samples
- Supplementary Table 3: Annotation of genomes by consortium
- Supplementary Table 4: Annotation of consortia, mean relative abundance
- Supplementary Table 5: Relative abundance of all consortia across all samples

**Supplementary Figures**

- Supplementary Figure 1: Kendall’s tau within the most frequent taxonomic groups at various levels
Supplementary Data

- Supplementary Data 1: Interactive display showing ordination of CAGs with taxonomic annotations
- Supplementary Data 2: gig-map displays for all consortia
- Supplementary Data 3: Relative abundance displays over time and across studies for all consortia
- Supplementary Data 4: Interactive UMAP of samples by consortium abundance
- Supplementary Data 5: Sequences of genes in all CAGs

Declarations

Ethics approval and consent to participate: The study was approved by the ethical committee of the Chinese Center for Disease Control and Prevention.

Consent for publication: Not applicable

Availability of data and materials: The data produced by this study has been made available in the Cirro Data Platform (https://cirro.bio) using anonymous login information – username: infant_microbiome_2023@cirro.bio; password: Public_Manuscript_Data123. To ensure the privacy of all participants, all sequences from the metagenomic data which align to the human genome have been masked from the FASTQ files which are provided publicly. Note that Supplementary Data 5 is only available via the Cirro platform due to the large file size (870MB).

The datasets supporting the conclusions of this article are included within the article and its additional files.
Competing interests: The authors declare that they have no competing interests.

Funding: SM was supported by funding from the Microbiome Research Initiative (Fred Hutch Cancer Center, PI: David Fredricks M.D.). SM, KMB, AFG, AJ, and JK were supported by funding from NIH NIAID R01AI127100.

Authors’ contributions: JK, XS, PB, and SS conceived and designed the study for the newly collected cohort. XS, LL, and YJ analyzed and generated data from physical specimens. SM performed bioinformatics analysis and implemented statistical analyses. SM, KMB, AFG, AJ, and JK collaboratively developed the statistical analysis approach. SM, PB, KMB, AFG, and JK collaboratively wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements: We would like to acknowledge the significant contribution of all study participants.

References


Bacterial Transmission during the First Few Months of Life. *Cell Host Microbe* 2018, 24:146-154 e144.


mechanism for transmaternal asthma protection induced by microbes. J Allergy Clin Immunol 2011, 128:618-625 e611-617.


