Gut DNA viromes of Malawian twins discordant for severe acute malnutrition

Alejandro Reyes[a,b,c,d], Laura V. Blanton[a,b], Song Cao[c], Guoyan Zhao[c], Mark Manary[e,f], Indi Trehan[e,g], Michelle I. Smith[a], David Wang[c,h], Herbert W. Virgin[c], Forest Rohwer[i], and Jeffrey I. Gordon[a,b,c,1]

[a]Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St. Louis, MO 63108; [b]Center for Gut Microbiome and Nutrition Research, Washington University School of Medicine, St. Louis, MO 63108; [c]Department of Pathology and Immunology, Washington University School of Medicine, St. Louis, MO 63110; [d]Department of Biological Sciences, Universidad de los Andes, Bogota, 111711 Colombia; [e]Department of Pediatrics, Washington University School of Medicine, St. Louis, MO 63110; [f]Department of Community Health, University of Malawi, Blantyre, Malawi; [g]Department of Pediatrics and Child Healthy, University of Malawi, Blantyre, Malawi; [h]Department of Molecular Microbiology, Washington University School of Medicine, St. Louis, MO 63110; and [i]Department of Biology, San Diego State University, San Diego, CA 92182

Contributed by Jeffrey I. Gordon, July 23, 2015 (sent for review June 5, 2015; reviewed by Frederic D. Bushman and David A. Relman)

The bacterial component of the human gut microbiota undergoes a definable program of postnatal development. Evidence is accumulating that this program is disrupted in children with severe acute malnutrition (SAM) and that their persistent gut microbiota immaturity, which is not durably repaired with current ready-to-use therapeutic food (RUTF) interventions, is causally related to disease pathogenesis. To further characterize gut microbial community development in healthy versus malnourished infants/children, we performed a time-series metagenomic study of DNA isolated from virus-like particles (VLPs) recovered from fecal samples collected during the first 30 mo of postnatal life from eight pairs of mono- and dizygotic Malawian twins concordant for healthy growth and 12 twin pairs discordant for SAM. Both members of discordant pairs were sampled just before, during, and after treatment with a peanut-based RUTF. Using Random Forests and a dataset of 17,676 viral contigs assembled from shotgun sequencing reads of VLP DNAs, we identified viruses that distinguish different stages in the assembly of the gut microbiota in the concordant healthy twin pairs. This developmental program is impaired in both members of SAM discordant pairs and not repaired with RUTF. Phage plus members of the Anelloviridae and Circoviridae families of eukaryotic viruses discriminate discordant from concordant healthy pairs. These results disclose that apparently healthy cotwins in discordant pairs have viromes associated with, although not necessarily mediators, of SAM; as such, they provide a human model for delineating normal versus perturbed postnatal acquisition and retention of the gut microbiota's viral component in populations at risk for malnutrition.

assembly of the human gut DNA virome | childhood malnutrition | age/disease-discriminatory phage and eukaryotic viruses | gnotobiotic mice | epidemiology

Malnutrition (undernutrition) is a leading cause of child mortality worldwide (1). Severe acute malnutrition (SAM) can manifest itself as progressive wasting (marasmus) or as a more abrupt onset syndrome characterized by generalized edema, hepatic steatosis, skin rashes and ulcerations, and anorexia (kwashiorkor). The configuration of the bacterial component of the gut microbiota of healthy infants evolves to an adult-like configuration during the first 2–3 y of life (2, 3). Normal postnatal maturation of the gut microbial community is perturbed in SAM; children with SAM living in Malawi and in Bangladesh have gut microbiota with bacterial configurations that appear younger (more immature) than the microbiota of chronologically age-matched individuals with healthy growth phenotypes (3, 4). Moreover, this immaturity is only transiently improved with current ready-to-use therapeutic food (RUTF) interventions (3, 4). These children can be viewed as having a persistent developmental abnormality—one that affects a microbial "organ" whose key functions include the biosynthesis of vitamins and the biotransformation of dietary components into products that benefit members of the gut microbial community and their host (2–5).

A study of 317 twin pairs from five rural villages in southern Malawi showed that discordance for moderate acute malnutrition (MAM) and SAM was surprisingly high during the first 3 y of life (43% of pairs) and not significantly different between mono- and dizygotic pairs (concordant undernourished pairs comprised 7% of the cohort) (4). The standard of care in Malawi is to treat both cotwins in pairs discordant for marasmus or kwashiorkor with a peanut-based RUTF for several weeks until a threshold increase in weight has been achieved (both siblings in the pair are treated to avoid potential problems arising from maternal food-sharing practices that emphasize the diseased child and neglect the healthy cotwin) (4, 6). Although short-term administration of RUTF has dramatically reduced mortality, it generally does not ameliorate the long-term morbidities associated with malnutrition—stunting, neurodevelopmental abnormalities, and immune dysfunction (e.g., refs. 6–10).

Transplantation of fecal samples obtained from children with kwashiorkor and their apparently healthy cotwins into separate groups of adult germ-free mice consuming a prototypic macro- and micronutrient-deficient Malawian diet resulted in transmission of

Significance

Childhood malnutrition is a global health problem not attributable to food insecurity alone. Sequencing DNA viruses present in fecal microbiota serially sampled from 0- to 3-y-old Malawian twin pairs, we identify age-discriminatory viruses that define a "program" of assembly of phage and eukaryotic components of the gut "virome" within and across pairs where both cotwins manifest healthy growth. This program is perturbed (delayed) in both members of discordant pairs where one cotwin develops severe acute malnutrition and the other appears healthy by anthropometry. This developmental delay is not repaired by therapeutic foods. These age- and disease-discriminatory viruses may help define familial risk for childhood malnutrition and provide a viral dimension for characterizing the developmental biology of our gut microbial "organ."

Author contributions: A.R. and J.I.G. designed research; A.R., L.V.B., M.M., I.T., and M.I.S. performed research; A.R. contributed new reagents/analytic tools; A.R., L.V.B., S.C., G.Z., D.W., H.W.V., F.R., and J.I.G. analyzed data; and A.R. and J.I.G. wrote the paper.

Reviewers: F.D.B., Perelman School of Medicine at the University of Pennsylvania; and D.A.R., Stanford University.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

Data deposition: The 16S rRNA and shotgun sequencing datasets have been deposited in the European Nucleotide Archive (ENA; www.ebi.ac.uk/ena) in raw format, prior to post-processing and data analysis, under study accession number PRJEB9818. The dataset of 17,676 viral contigs assembled from shotgun sequencing reads has also been deposited in ENA under the same accession number.

[1]To whom correspondence should be addressed. Email: jgordon@wustl.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1514285112/-/DCSupplemental.

MICROBIOLOGY

discordant weight loss and metabolic and gut barrier dysfunction phenotypes to the animals. Development of these pathologic phenotypes was diet-dependent: they were not observed, or dramatically mitigated, when gnotobiotic mice harboring a kwashiorkor microbiota received a healthy diet with adequate nutrients (4, 11). Together, these findings indicate that the gut microbiota is causally related to SAM (4) but also raise the question of how discordance for SAM arises and whether the cotwin classified as "healthy" by anthropometry has an underlying perturbation in his/her gut community that reflects familial risk for development of pathology. To address these issues, we focused on the most variable component of the human gut microbial community, the DNA virome. Moreover, although surveys of DNA viruses present in the gut microbiota of healthy adults had revealed a dominance of phage, in particular of lysogenic phages (prophages) (12–15), almost nothing was known about the normal pattern of assembly of the virome and the factors that shape this aspect of postnatal microbiota development (16).

Results

Fecal samples used for the present study had been collected from a subset of the larger 317 twin pair cohort: this subset consisted of 8 monozygotic and 12 dizygotic Malawian twin pairs between 0–30 mo of age living in five rural villages (see Table S1 for subject characteristics). Six of the discordant pairs contained a cotwin who developed kwashiorkor, whereas the other twin remained healthy. In the other six discordant pairs, one cotwin developed marasmus, whereas the other sibling remained healthy. Eight other pairs remained concordant for normal growth as defined by anthropometry (Table S1). Both siblings in each discordant pair were treated for 2–8 wk with a peanut-based RUTF. In addition, if fecal samples were available, we characterized the viromes of the twins' mother and an older sibling.

A total of 231 fecal samples were collected from twins, their mothers, and an older sibling at the time points shown in Fig. S1*A*. The samples were frozen immediately in cryogenic storage containers maintained at liquid nitrogen temperature, subsequently stored at −80 °C and then used as the starting material for purification of virus-like particles (VLPs) (*SI Methods*). VLP DNA isolated from each sample and technical replicates from six randomly selected samples were subjected to multiple displacement amplification (MDA) and shotgun pyrosequencing [53,334 ± 2,290 reads/sample (average ± SEM); 365 ± 121 nt/read (average ± SD); Table S1]. On average, 62.4 ± 23% (mean ± SD) of the reads per sample had no significant similarity to sequences in public DNA sequence databases (Fig. S1*B*) and 35 ± 23% (mean ± SD) had significant hits to an updated viral nonredundant (NR) database [Viral_NR_DB (13)], whereas only 0.95 ± 1.2% of reads had unique hits to a database of 128 human gut-associated bacterial genomes (17) (this latter result also highlights the quality of VLP purification before DNA extraction).

The dataset of raw pyrosequencing reads and a cross-assembly strategy described in Fig. S2*A* and *SI Methods* were used to assemble a total of 17,676 contigs ≥500 nt (largest, 228,572 nt); 85 ± 9% of the raw reads per VLP DNA sample mapped to these contigs when a threshold nucleotide sequence identity of ≥95% over the length of the read was applied (Table S1). Analyzing the size distribution of the contigs as a function of their sequencing coverage (Fig. S2*B*), and considering those with overlapping termini, we identified three distinct size ranges for circular contigs: (*i*) >30Kb (the expected size range for circular dsDNA phages belonging to the Caudovirales); (*ii*) 6–7 Kb [expected size for ssDNA phages in the Microviridae family, notably the Alpavirinae (18)]; and (*iii*) 3–4 Kb (expected size for ssDNA eukaryotic viruses in the Anelloviridae family). The results were consistent with our taxonomic assignments (Fig. S2*C* and Table S2): (*i*) 4,048 contigs had significant similarity to known members of the Caudovirales; (*ii*) 395 contigs were assigned to Microviridae [164 of these contigs were classified as belonging to the Alpavirinae, a recently described subfamily of temperate phages associated with members of the Bacteroidetes (18)]; and (*iii*) 2,414 contigs had significant similarity to members of the Anelloviridae, a family of single-stranded viruses that infect different eukaryotic hosts including humans (see Fig. S2 *D* and *E* for further taxonomic classification).

Features of Virome Assembly/Development in Young Malawian Twin Pairs. Raw reads from each virome were mapped to the assembled contigs and a normalized matrix of the number of reads per Kbp of contig sequence per million raw reads per VLP DNA sample (RPKM) was built (a "viral contig abundance matrix"; *SI Methods*). β-Diversity was measured using the Hellinger distance metric on the log-transformed matrix. This is analogous to using tables of bacterial operational taxonomic units (OTUs) for measurements of the degree of similarity between different (gut) microbial community samples. Distances were computed between fecal viromes sampled from a given individual over time (intrapersonal variation), as well as between individuals belonging to his/her family (interpersonal comparisons of cotwins, twin-mother or twin-older sibling), or to other families (interpersonal comparisons of unrelated twins, unrelated mothers/older siblings, or unrelated twins to unrelated mothers/older siblings).

The highest similarity between fecal DNA viromes was within an individual over time. Cotwins were more similar to each other; this relationship was not significantly affected by zygosity ($P = 0.47$; Mann–Whitney test); significantly greater differences in viromes were observed between a twin and his/her mother or older sibling or between any two unrelated individuals (Fig. S3*A*). The similarities between the fecal DNA viromes of unrelated young Malawian twins were significantly higher than between the twins and their mothers or older siblings (Fig. S3*A*). This latter finding emphasizes the importance of age as a variable affecting virome composition during the first 3 y of life. Age is also a major driver of variation in bacterial community composition, as illustrated by applying a phylogenetic distance metric (unweighted UniFrac) to 16S rRNA datasets generated from the same fecal samples used to purify VLPs (Fig. S3*B*).

Previous viromes sampled from fecal VLPs purified from healthy adult twin pairs and their mothers living in the United States showed that each individual's collection of viruses was highly distinctive and stable (13) (i.e., the Hellinger distances between viromes sampled from cotwins was not significantly different from the distances between the cotwin and mother or another unrelated person). Although a few phages were identified as shared across members of the small cohort of adult US twins (13, 19), the results emphasized the high degree of interpersonal variation that existed between these adults. In contrast, the Malawian twins, which differ from the adult US cohort in a number of respects including age, geographic location, health status, and hygiene practices, exhibited much more substantial similarity in their early-life DNA virome membership.

Age-Discriminatory Viral Contigs. To determine which viruses were responsible for the age signal described above, we first used a rarefied cross-assembly matrix to calculate two metrics; Shannon diversity and "predicted observed species." Samples were clustered into 5-mo age bins [a window wide enough to incorporate a sufficient number of samples for analysis while still being narrow enough to not compromise the number of time points (bins) needed to show age-associated changes]. Using this approach, both measures of α-diversity increase significantly as a function of age for single-stranded phages [0.129; $R^2 = 0.099$; $P < 0.0001$ (one-way ANOVA post test for linear trend); primarily members of Alpavirinae, which, as noted above, are associated with the Bacteroidetes (18)] and bacteria (slope for Shannon diversity index, 0.383; $R^2 = 0.501$; $P < 0.0001$). In contrast, the α-diversity of eukaryotic ssDNA viruses (Fig. S4) exhibits a negative albeit not statistically significant correlation with age (slope, −0.043; $R^2 = 0.015$; $P = 0.12$).

A Random Forests regression classifier was then trained to determine how well the chronologic age of a healthy donor of a given fecal sample could be predicted based on the DNA virome. Linear regression of predicted age against donor chronological age over the range of 6–22 mo yielded a regression coefficient of 0.6 (Fig. 1*A*). Fig. 1*B* shows a heat map of 22 age-discriminatory contigs that, when applying the Random Forests machine learning method, explain 68.7 ± 0.31% (mean ± SD) of the observed variance for concordant healthy twin pairs (compared with 54.5 ± 3.1% when using the full set of contigs) and yield a regression coefficient of 0.7. After summarizing the viral abundance matrix by the assignable taxonomy of its component contigs and sorting samples as a function of age, we found that (*i*) ssDNA eukaryotic viruses belonging to the Anelloviridae are highly abundant in the DNA viromes of healthy infants and children until 15–18 mo of age, after which time, the abundance of Anelloviridae diminish; (*ii*) ssDNA phages belonging to the Alpavirinae subfamily of the Microviridae increase in abundance as a function of age; and (*iii*) dsDNA phages assigned to the family Siphoviridae in the order Caudovirales are highly abundant in VLP samples from 0 to 10 mo of age and then slowly decrease (Fig. S5*A*). The high representation of this family from the Caudovirales during early phases of virome assembly and the high abundance of unclassified viruses during later months could reflect the high representation of phages from the Proteobacteria and Actinobacteria (early gut colonizers) in public DNA sequence databases and the paucity of full genomes for phages that use gut Bacteroidetes and Firmicutes as their hosts.

Viral Contigs That Distinguish Families. Table S3 shows that there were contigs present in at least one fecal VLP sample from all 20 families, as well as contigs present in up to 60% of the samples regardless of family. The observed family clustering (Fig. S6*A*) suggested that it should be possible to use viral contig representation to predict family of origin using Random Forests (see *SI Methods* and Fig. S7 for details about implementation including criteria used for discriminatory feature selection). The Random Forests classifier accurately assigned twin-pair DNA viromes by family of origin [Out-Of-Bag (OOB) error rate of 6.4 ± 0.66% (mean ± SD) using the discriminatory contigs]. A heat map of the abundances of the most discriminatory contigs revealed that twin pairs shared a large percentage of their viromes, whether they were concordant for healthy status or discordant for SAM (Fig. S6*B* and Table S4). Interestingly there were a significantly greater number of contigs that discriminate families with SAM discordance compared with those with concordant healthy twin pairs ($P = 0.02$; Kruskal–Wallis test; Fig. S6*C*).

Contigs That Distinguish Fecal DNA Viromes in Kwashiorkor and/or Marasmus Twin Pairs Compared with Concordant Healthy Pairs. β-Diversity measurements, based on the Hellinger metric, revealed that fecal viromes sampled from both members of SAM discordant pairs were less variable than those from concordant healthy pairs (i.e., both members of the discordant pairs did not develop more individualistic viromes). This significant reduction in variation was evident in kwashiorkor or marasmus discordant pairs at the time of diagnosis and endured during and following treatment with RUTF (Fig. 2*A*). These findings suggested that the DNA virome is perturbed in both members of these twin pairs, even though only one member manifests overt disease.

After removing family- and age-discriminatory contigs, we used Random Forests to determine whether health status could be predicted from virome composition. Classification proved to be quite accurate; 125 of the most discriminatory disease-associated viral contigs produced an OOB error rate of 9.61 ± 0.72% (see Fig. S5*B* for a heat map of these contigs as a function of health status). Importantly, their presence was not limited to the affected cotwin but rather was indicative of twin pairs discordant for marasmus or kwashiorkor; 80 of these disease-discriminatory contigs

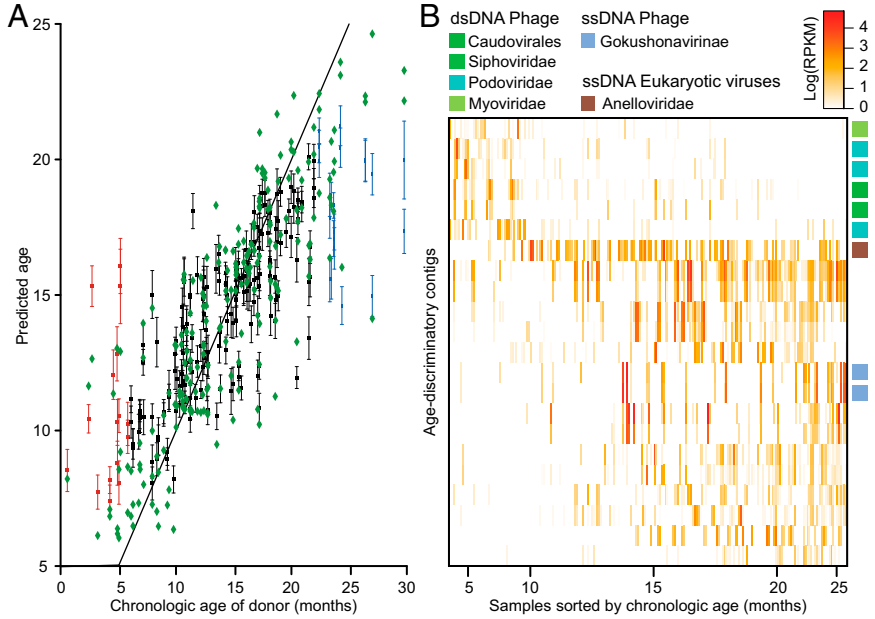

Fig. 1. Identification of age-discriminatory viral contigs. (*A*) Random Forests was used for a regression analysis to test if relative abundance of viral contigs is a good predictor of the human fecal microbiota donor's chronologic age. The dataset of assembled contigs from all viromes sampled from twin pairs was filtered to remove family-specific contigs (*SI Methods*). The percentage variation explained by the regression in 100 independent runs of the Random Forests algorithm was 54.5 ± 3.1% (mean ± SD). Predicted age (mean ± SD) for each fecal virome sample is plotted against the donor's chronologic age. Most errors in classification occur in samples obtained from donors <6 mo (red) and >23 mo of age (blue). Green diamonds indicate predicted age when using a sparse set of 22 of the most discriminatory contigs shown in *B*. The black diagonal represents the identity line ($y = x$). (*B*) Heat map of the abundance distribution of significantly discriminatory contigs as a function of age (months). Each row represents a significant age-associated viral contig. Boxes on the right are colored according to the contig's taxonomic annotation.

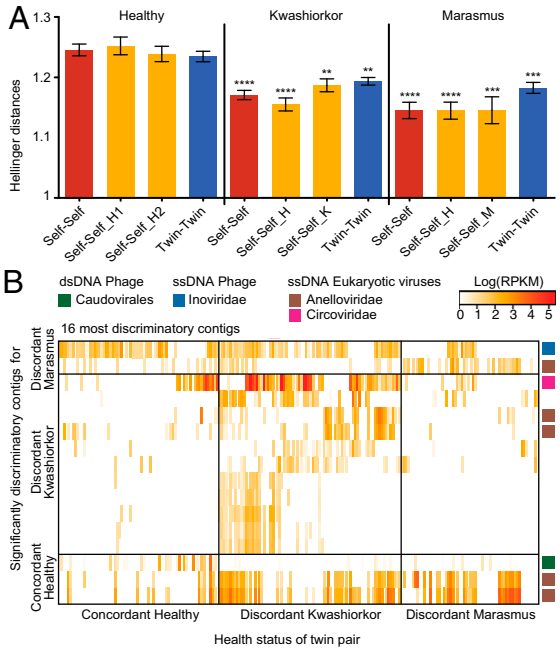

Fig. 2. Virome features associated with severe acute malnutrition. (*A*) The fecal viromes of twin pairs where one member developed SAM have reduced β-diversity compared with concordant healthy pairs. Pairwise Hellinger distances were calculated from the log-transformed viral contig abundance matrix. Comparisons within an individual over time (Self-Self) or between cotwins from the same family (Twin-Twin) are shown. Self-Self_H, healthy cotwin; Self-Self_K, kwashiorkor cotwin; Self-Self_M, marasmus cotwin. Each of the indicated comparisons (self–self, cotwin–cotwin, etc.) was referenced to the corresponding comparisons in concordant healthy pairs. (*B*) Contigs significantly associated with health status. Random Forests was used to classify samples. Each row represents a discriminatory contig and each column represents a VLP sample from a cotwin sorted by health status. The 16 contigs that best discriminate twin pairs discordant for marasmus and/or kwashiorkor from concordant healthy pairs are shown together with their assigned taxonomy.

were significantly associated with pairs containing a child with kwashiorkor, and another 18 contigs were significantly associated with pairs containing a cotwin with marasmus, whereas a third subset of 27 contigs were discriminatory for concordant healthy individuals, either based on their presence in these pairs or by their absence in healthy and presence in discordant pairs (see Table S4 for contigs that fall into these different categories and their annotations).

Fig. 2*B* presents a subset of 16 contigs, from the group of 125, which comprise a sparse Random Forests-derived model with an OOB error rate equivalent to that achieved with the full dataset of contigs. As is typical with viromes, most of the ORFs do not have significant similarity with known genes. The most common recognizable functions of their encoded proteins are related to virion structures (e.g., tail fibers, capsids) and integration of phage into the host genome (e.g., integrases, transposases, and regulatory genes such as *CI*). ORFs assigned to the Torque Teno Virus family were found in 5 of the 16 contigs. One of the disease-discriminatory contigs includes a gene specifying an Ig domain-containing protein similar to those identified by Minot et al. (15) and hypothesized to be responsible for a microbiome-derived adaptive immune system (20). We subsequently selected all proteins from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database with annotations of "Ig-like" or "Ig-domain" and blasted all identified proteins in the full dataset of 17,676 contigs to this Ig-only database (*e* value threshold, $1 \times e^{-3}$). A total of 384 proteins from 327 contigs had significant hits. The distribution of the Ig-like and IgA-domain proteins was compared among the taxonomically annotated contigs and contigs that were discriminatory for the different variables analyzed. Genes encoding these proteins were significantly enriched in the Caudovirales, in particular Podoviridae (χ^2 $P < 0.0001$). The presence of Ig-similar containing contigs was enriched in family-discriminatory but not health status-discriminatory contigs (χ^2 $P <$ 0.0001 and 0.636, respectively). A more comprehensive analysis of 70,160 predicted ORFs in all 17,676 contigs yielded 8,239 (11.74%) known or predicted proteins of which 64% were hypothetical or conserved hypothetical proteins; 44% of the remaining proteins were assigned to two KEGG categories "Nucleotide metabolism" and "Replication and repair," a predominance that is not surprising for viruses (13). The very small number of proteins in other KEGG categories precluded us from conducting a suitably powered analysis that tests whether there are any functions enriched in the subset of 22 age-discriminatory or 125 disease-discriminatory contigs we had identified.

There was no statistically significant effect of RUTF on the abundances of the diseased twin pair-discriminatory viral contigs compared with the last time point sampled before initiation of the food-based intervention ($P = 0.4822$; Friedman test) and no effect of zygosity (the latter conclusion comes with the caveat that the number of twin pairs studied is small). Moreover, these contigs were rare in older siblings or mothers with only one, belonging to Circoviridae, present at an abundance ≥ 1 RPKM in more than 50% of these other family members (Table S4*A*).

Thirty-seven of the disease pair-discriminatory contigs had assignable taxonomy to eukaryotic ssDNA viruses belonging to the Anelloviridae or Circoviridae; this number is significantly more than expected from the distribution of all contigs with assignable taxonomies (Fig. S6*D*; two-way ANOVA, $P = 0.0208$). (Note that our study revealed a large number of previously unidentified lineages in these two eukaryotic viral families; Fig. S2 *D* and *E* and *SI Methods*.)

Gnotobiotic Mouse Studies of the Kwashiorkor and Marasmus Viromes. The identification of eukaryotic ssDNA viruses as a prominent component of the set of viruses that discriminate SAM discordant twin pairs from concordant healthy pairs, but not the cotwin with normal anthropometry from his/her undernourished twin, raised the question of whether these viruses are causally related to disease pathogenesis. As noted above, transplantation of fecal microbiota from members of twin pairs discordant for kwashiorkor to adult germ-free mice consuming a prototypic Malawian diet transmitted discordant weight loss and metabolic phenotypes (including inhibition of the TCA cycle) and an enteropathy characterized by disruption of the small intestinal and colonic epithelial barrier (4, 11). Therefore, we isolated VLPs from fecal pellets that had been collected (and stored at −80 °C) from four groups of gnotobiotic mice sampled 3–32 d after the mice were gavaged with fecal microbiota obtained from kwashiorkor discordant pairs 56 and 57 [$n = 10$ animals/donor microbiota; donor microbiota were collected before subjects were treated with RUTF (4); see Table S5 for information about the time points when fecal samples were collected from mice; note that VLPs were also recovered from cecal contents harvested at the time the animals were killed]. Shotgun reads from the VLP DNAs [21,721 ± 6,869 reads/sample (mean ± SEM)] were used to query our dataset of 17,676 viral contigs from the human samples. A wide range (7–95%) of the reads from the different VLP samples mapped to a total of 87 contigs (successful transfer of these viruses to mice was defined as a contig present in at least one gut VLP sample at more than 0.1% relative abundance, with ≥10 reads mapping to that contig). Of these 87 contigs, 22 had assignable taxonomy; 19 of these belonged to the Caudovirales order (Table S5). Only one of the phages that were successfully transferred and retained represented a SAM discriminatory biomarker with assigned taxonomy—this single-stranded phage was classified as a member of the family Inoviridae (Table S5). The phages that were detected in recipient gnotobiotic mice

could either represent lytic viruses transferred with their corresponding host bacterial strains or prophages induced at various time points during the mouse experiment. No assignable eukaryotic viruses were detected in any of the fecal samples obtained from any of the mice at any of the time points surveyed, indicating that these human viruses were not retained in the guts of the gnotobiotic animals under the experimental conditions used. Our failure to capture these eukaryotic DNA viruses in mice is consistent with the fact that although a large number of Anelloviruses have been identified in domesticated and wild animals, including rodents, pigs, and nonhuman primates, successful infection of animal models using human Anelloviruses has yet to be reported (21).

Our findings also suggest that the Anelloviridae (and Circoviridae) detected in the microbiota of the SAM discordant pairs are not necessary or sufficient to produce the transmissible discordant weight loss, barrier disruption, and metabolic phenotypes previously documented in these recipient gnotobiotic mice. Complex transkingdom interactions are being documented between persistent enteric viruses (both DNA and RNA), members of the domain Bacteria as well as Eukarya, and components of the immune system (22–24). Anelloviruses have been identified as chronic infecting viruses, and have been isolated from multiple body habitats and biofluids including bile, feces, saliva, urine, amniotic fluid, breast milk, cervical secretions plus sewage, suggesting several potential routes of transmission (25, 26). Although early "infection" with members of this family is almost universal [100% within the first 2 y of life in one Japanese study (27), with constant shedding in feces during the first year documented in another study (28)], there is, as of yet, no proof that Anelloviruses cause any disease (29, 30). Changes in the abundances of Anelloviruses in serum have been reported in lung transplantation patients (31, 32), patients with diverse respiratory tract infections (33), and those who develop AIDS (34). SAM is also associated with defects in immune function, including disturbances in the gut mucosal barrier (9, 11). At present, it is not clear whether these viruses "simply" provide a high-resolution map of disordered immune regulation or whether they are mediators of various aspects of immune function and dysfunction. Addressing this question will be difficult; the host species specificities of Anelloviridae and Circoviridae and the inability to capture these viruses in gnotobiotic mice harboring human gut microbial communities (but not human immune cell repertoires) represent significant challenges to overcome when designing preclinical models for proof-of-concept tests of whether viruses in these families are "simply" biomarkers of SAM in these twin pairs or causally related to disease pathogenesis.

Epidemiologic Considerations. Because Random Forests proved successful in accurately classifying viromes based on age and health status, we attempted to use this machine-learning approach to classify viromes based on seasonality and/or village of origin. Accurate predictions were limited to village-of-origin (OOB error rate, 26.35 ± 2.4%). Using a subset of 162 village-discriminatory contigs, it was possible to decrease the OOB error rate to 14.89 ± 0.65% (mean ± SD); 105 of these contigs were members of the Anelloviridae and Circoviridae (Figs. S2 *D* and *E*, S6*D*, and S8*A*). The distributions of anthropometric measures [weight-to-height Z score (WHZ), weight-to-age Z score (WAZ), and height-to-age Z score (HAZ)] varied between the villages such that there were significant differences between some pairwise comparisons (one-way ANOVA and post hoc Tukey's tests). The greatest distinction involved Mitondo and Makhwira, which had more infants with lower WHZ and WAZ scores compared with Chamba, Mayaka, and M'biza (Fig. S8*B*). Mitondo and Makhwira are the two villages positioned in the lower Shire River Valley, where ambient temperatures are higher and rate of childhood illnesses, particularly malaria, are considerably greater than in the rest of the country (Fig. S8*C*). The distinct distributions of these virotypes should prove useful for subsequent epidemiologic and anthropologic studies that seek to address questions about factors that might affect how these viruses are acquired and transmitted between individuals and the contributions of these viruses to health status.

Discussion

We have conducted a time series comparative metagenomic study of the fecal DNA viromes of twins concordant for healthy status during the first 3 y of life and twins who became discordant for kwashiorkor or marasmus, plus their mothers and older siblings. Our results provide a human model for delineating normal versus perturbed postnatal acquisition and retention of the gut microbiota's viral component in children at risk for and with manifest undernutrition.

Although read depth was modest in this study, and larger sequencing depth can help identify rare virotypes and hypervariability in viral genomes (15), we benefited from the relatively longer read lengths obtained with pyrosequencing and the cross-assembly strategy used to generate partial or complete viral genomes [23- ± 2-fold (mean ± SEM) sequence coverage of assembled viral contigs]. Remarkably, 95.8% of the 70,160 predicted ORFs identified in the 17,676 assembled viral contigs encode hypothetical proteins or conserved hypothetical proteins. This finding further emphasizes the importance of developing new approaches, such as combining Hidden Markov Models with machine learning methods, to identify proteins that are highly discriminatory markers of the environmental origin and/or taxonomic features of viromes and hence the focus of efforts to delineate their functions.

Studying the assembly of components of the gut microbiota within and across families, including those containing mono- or dizygotic twin pairs, provides a microbial view of human postnatal development. Machine-learning methods (Random Forests) have yielded sparse models composed of a limited number of highly indicative age-discriminatory bacterial strains that together form a signature for defining the normal developmental biology of this microbial organ (3, 5). The fact that time-dependent changes in the representation of these indicative bacterial strains was similar across biologically unrelated individuals living in distinct geographic areas [e.g., Malawi and Bangladesh (3)] suggests that a set of (still-to-be-defined) rules govern development/differentiation of this organ which is composed of multiple cell lineages (taxa). The present study provides an additional developmental perspective, one focused on the viral component of the gut community. Analogous to the approach used for the bacterial component, applying machine-learning methods to a viral abundance matrix where contigs' abundances that were quantified as a function of individual, twin-pair, time after birth, family membership, and health status yielded a set of age-discriminatory phage and eukaryotic viruses.

The identification of viral contigs that discriminate both the kwashiorkor (or marasmus) and apparently healthy cotwins in discordant pairs from members of age-matched concordant healthy pairs is noteworthy from a developmental biology perspective: it reveals that specification of a normal community "fate" is perturbed in both members of a discordant twin pair and that the shared virome features of their "healthy" sibling provide an operational definition of a sensitized, at-risk host/microbial community.

The ability to compare and contrast phenotypes transmitted by microbiota from healthy cotwins in discordant pairs and microbiota from concordant healthy pairs to recipient gnotobiotic mice as a function of various defined perturbations (dietary, manipulations of innate and adaptive immunity, and other characteristics of the gut mucosal barrier, enteropathogen load, host genotype, etc.) provides a way to identify factors, both autonomous

(to the community) and nonautonomous (derived from the environment surrounding the community), that control the developmental trajectory of the microbiota and the origins of discordant phenotypes within twin pairs.

Answering the question of how the shared pattern of assembly/maturation of the DNA virome noted across biologically unrelated healthy infants and children is related to the program of succession/assembly of bacterial components of their microbiota and the codevelopment of their gut mucosal immune system may ultimately provide ways for deliberately advancing microbiota maturation in those with SAM (e.g., by introducing phage from a healthy individual whose chronologic age is similar to or older than that of the state of microbiota maturation in a child with SAM to facilitate colonization of human gut bacterial lineages that are not well represented in their developmentally arrested microbiota). Studies involving deliberate introduction of purified human fecal VLP preparations into gnotobiotic mice colonized with a defined consortium of sequenced members of the human gut microbiota have shown that viral–bacterial dynamics in vivo are complex; the correlations between phage and bacterial strain abundances are not always obvious and involve negative correlations shifted in time as a result of a predator–prey dynamic, whereas prophages have linear positive correlations (35). As a consequence, answering this question promises to be challenging.

Methods

Subjects were recruited through health centers located in the Malawian villages of Makhwira, Mitondo, M'biza, Chamba, and Mayaka using procedures approved by the College of Medicine Research Ethics Committee of the University of Malawi and by the Human Research Protection Office of Washington University School of Medicine in St. Louis. All experiments involving mice were performed with protocols approved by the Washington University Animal Studies Committee (4). Procedures for sample collection, purification of VLPs, shotgun pyrosequencing of VLP-derived DNA, assembly and annotation of viral genomes, cross-contig comparisons, calculations of viral α- and β-diversity, Random Forests analysis, viral phylogenetic analysis, bacterial 16S rRNA gene amplification and amplicon sequencing, as well as statistical analyses are described in detail in *SI Methods*.

ACKNOWLEDGMENTS. We thank Sabrina Wagoner, Su Deng, Jessica Hoisington-López, and Marty Meier for superb technical assistance. This work was supported by a grant from the Bill & Melinda Gates Foundation. L.V.B. received support from National Institutes of Health Grant T32 AI007172.

1. Black RE, et al.; Maternal and Child Nutrition Study Group (2013) Maternal and child undernutrition and overweight in low-income and middle-income countries. *Lancet* 382(9890):427–451.
2. Yatsunenko T, et al. (2012) Human gut microbiome viewed across age and geography. *Nature* 486(7402):222–227.
3. Subramanian S, et al. (2014) Persistent gut microbiota immaturity in malnourished Bangladeshi children. *Nature* 510(7505):417–421.
4. Smith MI, et al. (2013) Gut microbiomes of Malawian twin pairs discordant for kwashiorkor. *Science* 339(6119):548–554.
5. Subramanian S, et al. (2015) Cultivating healthy growth and nutrition through the gut microbiota. *Cell* 161(1):36–48.
6. Trehan I, Manary MJ (2015) Management of severe acute malnutrition in low-income and middle-income countries. *Arch Dis Child* 100(3):283–287.
7. Victoria JG, Kapoor A, Dupuis K, Schnurr DP, Delwart EL (2008) Rapid identification of known and new RNA viruses from animal tissues. *PLoS Pathog* 4(9):e1000163.
8. Gaayeb L, et al. (2014) Effects of malnutrition on children's immunity to bacterial antigens in Northern Senegal. *Am J Trop Med Hyg* 90(3):566–573.
9. Kosek M, et al.; MAL-ED network (2013) Fecal markers of intestinal inflammation and permeability associated with the subsequent acquisition of linear growth deficits in infants. *Am J Trop Med Hyg* 88(2):390–396.
10. Waber DP, et al. (2014) Impaired IQ and academic skills in adults who experienced moderate to severe infantile malnutrition: A 40-year study. *Nutr Neurosci* 17(2): 58–64.
11. Kau AL, et al. (2015) Functional characterization of IgA-targeted bacterial taxa from undernourished Malawian children that produce diet-dependent enteropathy. *Sci Transl Med* 7(276):276ra24.
12. Breitbart M, et al. (2003) Metagenomic analyses of an uncultured viral community from human feces. *J Bacteriol* 185(20):6220–6223.
13. Reyes A, et al. (2010) Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature* 466(7304):334–338.
14. Minot S, et al. (2011) The human gut virome: Inter-individual variation and dynamic response to diet. *Genome Res* 21(10):1616–1625.
15. Minot S, Grunberg S, Wu GD, Lewis JD, Bushman FD (2012) Hypervariable loci in the human gut virome. *Proc Natl Acad Sci USA* 109(10):3962–3966.
16. Breitbart M, et al. (2008) Viral diversity and dynamics in an infant gut. *Res Microbiol* 159(5):367–373.
17. Forsberg KJ, et al. (2012) The shared antibiotic resistome of soil bacteria and human pathogens. *Science* 337(6098):1107–1111.
18. Krupovic M, Forterre P (2011) Microviridae goes temperate: Microvirus-related proviruses reside in the genomes of Bacteroidetes. *PLoS One* 6(5):e19893.
19. Dutilh BE, et al. (2014) A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat Commun* 5:4498.
20. Barr JJ, et al. (2013) Bacteriophage adhering to mucus provide a non-host-derived immunity. *Proc Natl Acad Sci USA* 110(26):10771–10776.
21. Nishiyama S, et al. (2014) Identification of novel anelloviruses with broad diversity in UK rodents. *J Gen Virol* 95(Pt 7):1544–1553.
22. Reese TA, et al. (2014) Coinfection. Helminth infection reactivates latent γ-herpesvirus via cytokine competition at a viral promoter. *Science* 345(6196):573–577.
23. Baldridge MT, et al. (2015) Commensal microbes and interferon-λ determine persistence of enteric murine norovirus infection. *Science* 347(6219):266–269.
24. Nice TJ, et al. (2015) Interferon-λ cures persistent murine norovirus infection in the absence of adaptive immunity. *Science* 347(6219):269–273.
25. Breitbart M, Rohwer F (2005) Method for discovering novel DNA viruses in blood using viral particle selection and shotgun sequencing. *Biotechniques* 39(5):729–736.
26. Bernardin F, Operskalski E, Busch M, Delwart E (2010) Transfusion transmission of highly prevalent commensal human viruses. *Transfusion* 50(11):2474–2483.
27. Ninomiya M, Takahashi M, Nishizawa T, Shimosegawa T, Okamoto H (2008) Development of PCR assays with nested primers specific for differential detection of three human anelloviruses and early acquisition of dual or triple infection during infancy. *J Clin Microbiol* 46(2):507–514.
28. Kapusinszky B, Minor P, Delwart E (2012) Nearly constant shedding of diverse enteric viruses by two healthy infants. *J Clin Microbiol* 50(11):3427–3434.
29. Virgin HW, Wherry EJ, Ahmed R (2009) Redefining chronic viral infection. *Cell* 138(1): 30–50.
30. Okamoto H (2009) History of discoveries and pathogenicity of TT viruses. *Curr Top Microbiol Immunol* 331:1–20.
31. De Vlaminck I, et al. (2013) Temporal response of the human virome to immunosuppression and antiviral therapy. *Cell* 155(5):1178–1187.
32. Young JC, et al. (2015) Viral metagenomics reveal blooms of anelloviruses in the respiratory tract of lung transplant recipients. *Am J Transplant* 15(1):200–209.
33. Maggi F, et al. (2003) TT virus in the nasal secretions of children with acute respiratory diseases: Relations to viremia and disease severity. *J Virol* 77(4):2418–2425.
34. Thom K, Petrik J (2007) Progression towards AIDS leads to increased Torque teno virus and Torque teno minivirus titers in tissues of HIV infected individuals. *J Med Virol* 79(1):1–7.
35. Reyes A, Wu M, McNulty NP, Rohwer FL, Gordon JI (2013) Gnotobiotic mouse model of phage-bacterial host dynamics in the human gut. *Proc Natl Acad Sci USA* 110(50): 20236–20241.
36. Manary MJ (2006) Local production and provision of ready-to-use therapeutic food (RUTF) spread for the treatment of severe childhood malnutrition. *Food Nutr Bull* 27(3, Suppl):S83–S89.
37. Kleiner M, Hooper LV, Duerkop BA (2015) Evaluation of methods to purify virus-like particles for metagenomic sequencing of intestinal viromes. *BMC Genomics* 16:7.
38. Fouts DE (2006) Phage_Finder: Automated identification and classification of prophage regions in complete bacterial genome sequences. *Nucleic Acids Res* 34(20): 5839–5851.
39. Li W, Godzik A (2006) Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22(13):1658–1659.
40. Niu B, Zhu Z, Fu L, Wu S, Li W (2011) FR-HIT, a very fast program to recruit metagenomic reads to homologous reference genomes. *Bioinformatics* 27(12):1704–1705.
41. Caporaso JG, et al. (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 7(5):335–336.
42. Díaz-Uriarte R, Alvarez de Andrés S (2006) Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* 7:3.
43. Kursa MB (2014) Robustness of Random Forest-based gene selection methods. *BMC Bioinformatics* 15:8.
44. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22(22):4673–4680.
45. Tamura K, et al. (2011) MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* 28(10):2731–2739.
46. Page RD (1996) TreeView: An application to display phylogenetic trees on personal computers. *Comput Appl Biosci* 12(4):357–358.
47. Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* 73(16): 5261–5267.