Individuals at risk for developing rheumatoid arthritis harbor differential intestinal bacteriophage communities with distinct metabolic potential

Mihnea R. Mangalea[1], David Paez-Espino[2,3], Kristopher Kieft[4], Anushila Chatterjee[1], Jennifer A. Seifert[5], Marie L. Feser[5], M. Kristen Demoruelle[5], Meagan E. Chriswell[5], Alexandra Sakatos[3], Karthik Anantharaman[4], Kevin D. Deane[5], Kristine A. Kuhn[5], V. Michael Holers[5], and Breck A. Duerkop[1,6,*]

[1]Department of Immunology and Microbiology, University of Colorado School of Medicine, Aurora, CO, USA

[2]Mammoth Biosciences, San Francisco, CA, USA

[3]Anicilia Therapeutics, New York, NY, USA

[4]Department of Bacteriology, University of Wisconsin-Madison, Madison, WI, USA

[5]Division of Rheumatology, University of Colorado School of Medicine, Aurora, CO, USA

[6]Lead Contact

[*]Correspondence: breck.duerkop@cuanschutz.edu

SUMMARY

Rheumatoid arthritis (RA) is an autoimmune disease characterized in seropositive individuals by the presence of anti-cyclic citrullinated protein (CCP) antibodies. RA is linked to the intestinal microbiota, yet the association of microbes with CCP serology and their contribution to RA is unclear. We describe intestinal phage communities of individuals at risk for developing RA, with or without anti-CCP antibodies, whose first degree relatives have been diagnosed with RA. We show that at-risk individuals harbor intestinal phage compositions that diverge based on CCP serology, are dominated by Lachnospiraceae phages, and originate from disparate ecosystems. These phages encode unique repertoires of auxiliary metabolic genes (AMGs) which associate with anti-CCP status, suggesting that these phages directly influence the metabolic and immunomodulatory capability of the microbiota. This work sets the stage for the use of phages as preclinical biomarkers and provides insight into a possible microbial-based causation of RA disease development.

KEYWORDS

Bacteriophages, rheumatoid arthritis, autoimmune disease, microbiome, phage-host interaction, phage-host metabolism

INTRODUCTION

Rheumatoid arthritis (RA) is a systemic autoimmune disease with a global prevalence of approximately 1%. The development of RA in at-risk individuals is dependent on a combination of genetics, epidemiological factors, and systemic immune dysregulation [1]. The heritability of RA is estimated to be 40–60%, with increased familial risk evident among first-degree relatives (FDRs) of individuals with diagnosed RA [2, 3]. Analyses of at-risk FDRs, even those without serum RA-related autoantibodies, have identified patterns of mucosal inflammation whereby anti-cyclic citrullinated peptide (anti-CCP) antibodies and rheumatoid factors (RF), as well as cytokines and chemokines, are expressed locally in a subset of individuals [4-6]. In addition, anti-CCP and RF are present in the blood for years prior to the onset of RA, and their presence as well as circulating cytokine and chemokine biomarkers, are predictive of future RA development [7-9]. To probe the mucosal origins hypothesis [1] and the mounting evidence implicating intestinal microbiota perturbations in RA etiopathogenesis [10], it is necessary to characterize the ecological associations of the microbiota in at-risk individuals susceptible to RA.

Studies linking the role of the intestinal microbiota to systemic autoimmune diseases predominantly rely on 16S ribosomal gene analyses of bacteria within the microbiome, and have expanded our understanding of dysbiosis in the RA intestine. Individuals with established RA harbor a microbiota dominated by *Prevotella copri* [11, 12], enriched with Gram-positive bacteria [13], and decreased carriage of bifidobacteria [14], Gram-negative *Bacteroides,* and Firmicutes [13, 15]. The association of enriched Prevotellaceae, including *P. copri*, has also been described in individuals with preclinical RA [16], indicating that intestinal *P. copri* is immune-relevant to the pathogenesis of RA [17]. The presence of *P. copri* may therefore represent a biological indicator and additional risk factor for RA development and progression [18]. However, associating a single organism to RA etiology neglects the interactions of bacteria with their surrounding environment and other bacterial community members whose populations can be influenced by predatory bacteriophages (phages).

In contrast to the recent enthusiasm for characterizing microbial links to the etiology of RA, relatively little is known concerning the composition of phage communities in the intestine as it relates to RA disease risk. Phages of the intestinal microbiota can fluctuate in community composition in response to immune system function and disease, which suggests that they could be exploited as biomarkers for early disease detection [19]. Metagenomic sequencing strategies have revealed extensive and diverse populations of phages in the human intestine [20-22], in which phage community dynamics correlate with distinct disease states [23-25]. Specific intestinal phage genomic signatures precede autoimmunity development of type 1 diabetes in a cohort of diabetes-susceptible children, with disease-associated phages correlating to the bacterial component of the microbiota [26]. In addition to the direct impact of intestinal phages on bacterial community composition via classical predation and prophage mediated bacterial competition and metabolism, phages also adhere to mucosal surfaces, significantly impacting microbial colonization [27] and host mucosal immunity development [28]. Evidence is emerging that phages are also immunomodulatory through intrinsic anti-inflammatory properties, and are capable of direct lymphocyte regulation through the ability to translocate to multiple tissues and organs [29]. Despite these observations and potential implications for systemic autoimmune diseases like RA, evaluation of intestinal phages in the context of RA disease risk has yet to be described.

The interplay between intestinal bacteria, their phages, and the host immune system, whose interactions have consequences not only for compositional dysbiosis but immunomodulation, must be considered in the etiopathogenesis of RA. The microbiome, and more recently the virome, have been implicated in a range of human diseases including cancers [30, 31], inflammatory bowel diseases [32, 33], and arthritis [11, 34]. By characterizing the phage populations in an at-risk RA FDR cohort; further sub-grouped with regard to autoantibody status as defined by the presence of anti-CCP antibodies and compared to healthy controls, we have begun to address this question. The cohort contains individuals that do not have inflammatory arthritis or established RA disease but are FDRs to an individual with diagnosed RA, which alone increases RA risk. Studying the microbiomes of at-risk individuals in the preclinical RA state could lead to the identification of biomarkers and therapeutic targets independent of confounding by the use of drugs in subjects with active arthritis.

We used metagenomics to define intestinal phage populations of anti-CCP positive (CCP+) and negative (CCP-) individuals in an at-risk FDR cohort. Phage matching to bacterial hosts showed divergent intestinal phage communities dependent on anti-CCP serology status. We observed an overabundance of phages targeting Bacteroidaceae and Sreptococcaceae bacteria in CCP+ at-risk FDRs as well as phages targeting Bacteroidaceae bacteria in CCP- at-risk FDRs. Importantly, analysis of the metabolic traits encoded in phage metagenomes revealed intra-cohort profiles reflecting distinct immunomodulatory potential. Phages with auxiliary metabolic genes (AMGs) that modify lipopolysaccharide and other outer membrane glycans of host bacteria were differentially abundant, implicating modifications to bacterial antigenicity [35] and bacterial fitness [36] in RA-associated communities. Core phage metabolic genes, including 14 genes which are globally conserved among phages from multiple diverse environments [37], as well as bacterial surface modifying enzymes, were associated with phages targeting *Flavonifractor* sp. in the CCP+ cohort and *Bacteroides* sp. in the CCP- cohort. Phages targeting Lachnospiraceae (*Clostridium scindens*) and Actinomyces (*A. oris*), including several AMGs, were over-abundant among CCP+ and CCP- individuals, respectively, compared to healthy controls. Our data show that there are unique and abundant intestinal phages specific to RA-susceptibility status, and this highlights their potential as biomarkers for preclinical RA and the need for further pursuit of community-level bacteria-phage interactions during the development and progression of RA.

RESULTS

First-degree relatives to individuals with rheumatoid arthritis.

A total of 25 human subjects were identified from the Studies of the Etiology of Rheumatoid Arthritis (SERA) [38], including 16 FDRs of individuals with RA and 9 age and sex matched healthy controls (HC). FDR subjects for which a detectable level of anti-CCP autoantibody was present (defined by a value of $\geq$ 20 units/mL in either ELISA assay for anti-CCP3.1 IgA/IgG or anti-CCP3 IgG (Inova Diagnostics) [39]) were designated the CCP+ group (n = 8). FDRs with no anti-CCP detected were designated the CCP- group (n = 8) (Table 1). Mean ages for the three groups in this study were 61.3 ± 11.0 for CCP+, 49.0 ± 15.7 for CCP-, and 44.4 ± 13.6 for HC. The distribution of sexes for each group is reported as percent female, with 88.9% for CCP+, 62.5% for CCP-, and 66.7% for HC. Among the CCP+ and HC groups, 3/9 and 2/9 of individuals have reported ever smoking (a risk factor associated with RA), respectively (Table 1).

Generation and curation of *de novo* assembled VLP contigs.

We used individual fecal samples from the subjects obtained at the time of autoantibody and clinical evaluations, and isolated total genomic DNA for shotgun metagenomic sequencing using an untargeted amplification-independent approach [23, 40]. Samples were physically separated into whole metagenome (M), including all genomic DNA present in the sample, and virus-like particle (VLP) fractions, which were subjected to phage-specific precipitation (Figure S1A). Illumina sequencing resulted in an average of 123.8 ± 32.2, 135.2 ± 40.4, and 104.7 ± 45.9 million (M) paired end reads per sample for CCP+, CCP- and HC whole metagenomes, respectively, and an average of 67.3 ± 29.5, 73.2 ± 33.7, and 89.6 ± 47.8 M paired reads per sample for CCP+, CCP- and HC VLP fractions, respectively (Figure S1B). VLP sequencing reads were used for *de novo* contig assembly of VLP metagenomes. In total, 3.56 M contigs were assembled and pooled from the 25 individual metagenomes, with 80,762 contigs longer than 5 kb (Figure 1A). VLP contigs longer than 5 kb were distributed evenly across the three sample groups, totaling 2908.6 ± 1461.3, 3209.0 ± 2573.8, and 3535.7 ± 2826.4 contigs per sample for CCP+, CCP- and HC respectively (Figure S1C).

These 80,762 contigs served as a starting point for identifying putative phages using a three-pronged approach of independent phage discovery methods (Figures 1A and S1D). The first method (P/M ratio) employed a previously validated read mapping strategy whereby VLP read sets from all 25 samples were mapped to both whole metagenome (M) and VLP (P) contigs [23]. Using the read-mapping P/M ratio (see Methods), we identified 2,117 unique putative phage contigs after dereplication at 95% sequence identity. Next, we identified an independent set of phage contigs by aligning all open reading frames (ORFs) of the 80,762 VLP contigs against a set of 25,281 curated viral protein families (VPFs) [41]. Using this VPF method, several filters were applied to identify viral contigs; (i) 2,902 contigs were identified as having 5 or more VPF hits and non-viral Pfam hits below 20% of total ORFs on a contig, (ii) 263 contigs were identified with 5 or more VPF hits and less than 50% non-viral Pfam hits on a contig, (iii) 644 contigs with 2-4 VPF hits and 0 non-viral Pfams, (iv) 976 contigs with at least 1 VPF hit, without considering any non-viral Pfams. In total, after dereplication, the viral contigs arising from all above filters resulted in 4,785 unique viral contigs. For the third and final approach we employed VIBRANT (Virus Identification By iteRative ANnoTation), a sequence-independent algorithm that uses neural networks of viral protein signatures to identify lytic and lysogenic phages [37]. Using VIBRANT, we identified 4,758 unique viral contigs.

To consolidate this list, we identified contigs that were shared between all three phage discovery methods, resulting in a curated list of 660 contigs (Figures 1A and 1B). This curated list of putative phage contigs range in size from 5,007 bp to 557,525 bp. To assess host bacterial contamination among these contigs, we employed CheckV, a pipeline for assessing the quality of viral genomes [42]. CheckV analysis revealed a reduced level of host bacterial contamination and an increase of pure viral genomes in the final list of 660 curated contigs as compared to varying levels of contamination among the three separate methods prior to contig overlap identification (Figure 1C). We estimated completeness of our curated contigs using CheckV and determined a greater distribution of “high quality” contigs relative to contig length, in comparison to the three independent methods (Figure S2) [43]. Further, using the VIBRANT platform for integrated provirus prediction, we describe communities of predominantly lytic viral genomes belonging to Siphoviridae morphology (Figure S3). By using a combination of approaches for viral contig discovery and assessing the overlap among these methods, we have extracted a set of 660 predicted phages which are of overall high quality, both in terms of viral contig completeness and lack of bacterial contamination than those from each of the individual methods (Figures 1C and S2), which to date have been used primarily in isolation to identify and characterize viral metagenomes.

Clustering of metagenomic viral contigs reveals distinct viral ecological composition.

Next we compared our set of curated contigs to over 2.3 million viral whole genome and metagenome sequences from the IMG/VR database [44]. We used blastn at a threshold of 95% sequence identity over 85% of 1 kb sequence length and Markov clustering to group our contigs with related sequences from IMG/VR. Of the 660 contigs, 346 (52.4%) clustered into 255 clusters that contained 7,736 additional metagenomic viral contigs (mVCs) from IMG/VR. The remaining 314 contigs (47.6%) were classified as singletons, with an even distribution among CCP cohorts compared to healthy controls (Figure S4A). Of the curated contigs that were clustered, cluster sizes ranged from 2 to 646 members with 78.4% of the groups containing more than 2 partners and 36.5% containing more than 10 members, and 65.9% between 2 – 10 members (Figure S4B). Among these 255 clusters, 14 included reference prophages and lytic phages, and 318 (48.2%) clustered with classified mVCs, thus assigning multiple levels of taxonomy to our contigs (Figures 2A, 2B, and Supplementary Table 1).

Although host assignments were made using sequence-based clustering, host specificity was further determined by aligning Clustered Regularly Interspaced Short Palindromic Repeat (CRISPR) spacer sequences to our 660 curated contigs. CRISPR-Cas serves as a snapshot of previous phage infections in the form of acquired spacer sequences that represent invading viral genomes [45], and these sequences can be used for accurate identification of phage-host interactions in intestinal microbiomes [23, 46]. CRISPR spacer host assignments at the family level were present in 207 of 660 contigs (31.4%). All CRISPR spacer queries considered for these analyses, ranging in length from 18 to 70 bp, were matches of 93.1–100% identity across the full length of the query and allowing for 0–2 mismatches and up to 1 gap throughout [47] (Supplementary Table 2). Among predicted phages, total assigned CRISPR spacers were evenly distributed, yet CCP+ sample containing phages predicted to target Lachnospiraceae*,* Ruminococcaceae*,* Streptococcaceae, and Veillonellaceae bacterial families were disproportionately abundant (Figures 2A and 2B). In total, 21 bacterial families were identified as hosts via CRISPR spacer matching, supplementing the phage-host interactions discerned from sequence-based clustering (Figure 2A). Among all samples in this study, phages were predicted to target Lachnospiraceae, Ruminococcaceae, Clostridiaceae and Bacteroidaceae bacteria with highest frequency of total CRISPR spacers (Figure 2A). Phage-host interactions were also measured in terms of host range specificity, showing that while the majority of the phages were predicted to have narrow host ranges, several spacers were linked to multiple hosts across family level and higher taxa (Figure 2C), consistent with prior observations of diverse viromes [47]. Broad host range phages were found across all cohorts, but particularly among CCP+ sample contigs (Figure 2D) suggesting a more dysbiotic community of host bacteria among these individuals' metagenomes.

We further explored the association of sample cohorts to phage hosts using read mapping to determine differential host abundance profiles (Figure 3). Reads from all samples were mapped to assembled phage contigs whose host assignments were deduced using

CRISPR-spacer matching and Markov clustering to quantify sequence abundances by measuring cohort-based read recruitment [23, 48-50]. In comparing the differential read recruitment to phages predicted to infect separate bacterial families, we observed differences based on reads originating from either the CCP+ or CCP- groups in relation to the HC cohort (Figure 3). Among the most striking, phage contigs targeting Bacteroidaceae recruited significantly more reads from CCP+ viromes than either HC or CCP- individuals (Figure 3A). In contrast, phages predicted to target Clostridiaceae bacteria were evenly abundant across all three groups (Figure 3B). For Lachnospiraceae bacteria, CCP+ phages recruited were evenly distributed among the groups with a slight elevation in CCP+ individuals that was not statistically significant (Figure 3C). Ruminococcaceae phages were significantly skewed when comparing HC to CCP- individuals (Figure 3D) and a major shift in phage read recruitment abundance was evident for Streptococcaceae phages, as a greater percentage of total CCP+ reads were mapped to these phages in relation to either HC or CCP- virome reads (Figure 3E). This skew among CCP+ individuals is supported by prior works showing elevated Streptococcal phage abundances in intestinal viromes of humans with inflammatory bowel disease [32] and a murine model of colitis [23]. Lastly, no significant differences were observed for read recruitment to Veillonellaceae-targeting phages (Figure 3F). Thus, differences in the host specificities were evident between CCP+, CCP-, and HC groups with respect to read mapping abundance profiles for Bacteroidaceae, Ruminococcaceae, and Streptococcaceae phages.

CRISPR spacer host metadata reveal CCP+ phages represent greater variability in microbial host ecology.

To further explore the phage ecology from our subject cohort, we analyzed the host and mVC metadata from the Joint Genome Institute’s (JGI) Genomes OnLine Database (GOLD) [51]. The JGI GOLD database contains metadata from over 100,000 biosamples and over 350,000 sequencing projects involving genomic and metagenomic sequencing data from biological isolates worldwide. Moreover, recent work has contributed an additional 52,515 metagenome-assembled genomes from diverse ecologies and geographic distributions [52], further enhancing microbial host ecosystem analysis. Using the GOLD Biosample Ecosystem Classification system, we analyzed the ecosystem distributions for all CRISPR spacers identified in our curated contig list and discovered that the majority of host assigned contigs fell within four distinct ecosystem classification levels; from broad to specific environments: host-associated, human-associated, digestive system, and large intestine (Figure 4). For phages that were previously identified as having CRISPR spacer host assignments, total spacer alignments as identified by blastn ranging from 1 to 825 per contig, were tallied and used to calculate the uniformity of spacer origins per contig (Supplemental Table 3). For each of the four ecosystem categories, the most abundant classifications were used to compare across the study cohorts. At the highest order GOLD Ecosystem distribution, the host-associated (i.e., human, mammal, plant, arthropod, fungi) origin classification per contig was comparable for the HC and CCP- groups but not for the CCP+ group (Figure 4A). A similar pattern was evident at the lower order metadata distributions, with phage contigs derived from CCP+ individuals being more divergent from the other cohorts for contigs of human-associated origin (Figure 4B), digestive system origin (Figure 4C), and large intestine origin for the Ecosystem Subtype (Figure 4D).

These compositions of multiple CRISPR spacer ecosystem distributions reveal homogeneity among phages derived from HC and CCP- samples, and indicates more dysbiotic communities across CCP+ samples, suggesting that CCP+ individuals harbor disparate phage communities that are more likely to originate from non-host associated sources. The putative origins of these phages are related to environmental metadata of CRISPR spacers in the JGI GOLD database describing the origin of bacterial DNA samples across ecologically diverse biomes worldwide [52]; and increased heterogeneity in the CCP+ phages suggests a condition-dependent host intestinal environment that maintains diversity. At the highest Ecosystem classification level, with only three unique classification terms, these non-host associated sources that are more prevalent in the CCP+ group, correspond to a higher degree of spacers matching organisms originating from environmental and/or engineered habitats as archived in GOLD (Figure S5). The ecosystem distributions of Category, Type, and Subtype have 43, 126, and 146 unique terms for each classification level respectively, indicating multiple possible combinations for organism habitats. Thus, our analysis of GOLD metadata for all phages with predicted host isolates within our study reveals divergent habitat origins for CCP+ derived contigs.

Quantitative read mapping reveals differentially abundant contigs despite sample cohesiveness.

We next asked whether certain phage community members are present in different abundances among the members of the cohort at-risk for rheumatoid arthritis compared to healthy controls. To assess differences between phages among the sample groups, we used a viral read recruitment strategy whereby VLP reads from all samples were mapped to the 660 curated contigs [23, 48]. Using read count matrices for all contigs as input in the DEseq2 statistical package for differential analysis of comparative count data [53], we analyzed three pairwise comparisons for over- or under-abundant viral contigs (Figure 5). Initial comparisons of the normalized and log-transformed count matrices were performed to evaluate the experiment-wide trends across all samples. Principal component analyses reveal minimal variance explained by the first two principal components for CCP+ vs HC samples (Figure 5A), CCP- vs HC samples (Figure 5B), and CCP+ vs CCP- samples (Figure 5C), indicating that total sample community signatures cannot be readily differentiated based on at-risk or healthy control cohorts. We further explored the sample similarities by comparing Euclidian sample-to-sample distances of the regularized log-transformed count matrices. Hierarchical clustering of sample-to-sample distances did not reveal any discernable clustering for CCP+ vs HC samples (Figure 5D), and only minimal similarities between two CCP- samples when compared to the HC

(Figure 5E) and CCP+ (Figure 5F) groups, suggesting general sample cohesiveness between cohorts.

We next analyzed specific members of the intestinal phage community, considering the rationale that samples with complex communities are better explored at the level of each unique member [33]. Visualization of the principal components incorporating the viral identification metrics used in the VIBRANT neural network for our 660 curated contigs shows minimal differentiation among phage scaffolds based on scaffold quality (Figure S6A) or predicted phage state (i.e., lytic or lysogenic) (Figure S6B), although fragmentation of smaller sized contigs is evident for both analyses. Further, grouping of contigs at the sample type level does not differentiate any specific cluster (Figure S6C), which is consistent with the minimal variance observed at the sample level (Figures 5A, 5B, and 5C). Finally, we assessed the differential abundance of read recruitment counts for the set of 660 contigs and estimated fold changes based on the negative binomial generalized linear model provided by DESeq2 [53]. Using thresholds of log2-fold change greater than 1 or less than -1 (equivalent to fold change of ± 2) and Benjamini-Hochberg adjusted *p*-values < 0.001, we identified a total of 178 differentially abundant contigs (27% of the 660 phages) across three pair-wise abundance comparisons. For CCP+ vs HC samples a total of 59 contigs (30 over- and 29 under-abundant) (Figure 5G), for CCP- vs HC a total of 66 contigs (27 over- and 39 under-abundant) (Figure 5H), and for CCP+ vs CCP- a total of 53 contigs (27 over- and 21 under-abundant) (Figure 5I) passed our thresholds for significance. This suggests that there are unique changes in select phage abundances from the intestinal viromes of individuals at risk for RA, and that these changes are more nuanced than sample-based community associations can reveal. These data indicate that these cohort groups represent minimal sample-sample variation, but may provide clues related to detection of biomarkers via specific community members. The top phage contigs associated with either CCP+ or CCP- individuals were *Clostridium scindens* (Lachnospiraceae) and

Actinomyces oris (Actinomycetaceae), respectively, over-abundant at $\log_2$ fold changes of 25.9 and 23.5 compared to the healthy control samples.

A comparison of the bacterial relative abundances via 16S amplicon sequencing confirmed an expansion of Lachnospiraceae bacteria among samples originating from CCP+ individuals (Figure S7A). This confirms, in part, observations of over-abundant Lachnospiraceae-targeting phage contigs for the CCP+ but not CCP- cohorts (Figures 6B and 6C). The bacterial composition across all cohorts was relatively even in terms of richness (Figures S7B and S7C), evenness (Figure S7D), and species diversity (Figure S7E). Conversely, phage host abundances in the CCP- cohort relative to healthy controls were not correlated to a family-level differentiation in bacterial taxa relative abundance.

Phage auxiliary metabolic gene abundances highlight cohort-associated disparities in metabolic potential.

To determine the functional potential and metabolic capabilities within intestinal phages, we quantified AMGs assigned to specific metabolic pathways in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database across at-risk and healthy cohorts. Since their identification as viral drivers of host metabolism [54], phage-encoded AMGs have been recognized as consequential actors that redirect host functional capacities thereby directly influencing local ecology [55, 56]. Analyses of AMGs using VIBRANT and KEGG pathway annotations can provide valuable insights into potentially altered metabolic functions or informative biosignatures for cohort-associated microbial communities [37, 57]. To this end, we assessed our set of curated phage contigs against 2,835 AMGs with KEGG annotations identified as “metabolic pathways” or “sulfur relay system” [37]. Among our 660 phage contigs, 161 (24%) were found to encode at least 1 AMG, with 252 AMGs in total across all samples (Supplemental Table 4). Phages originating from the HC cohort accounted for 131 metabolic signatures, while CCP+ and CCP- had less total AMGs with 77 and 44, respectively (Figure S8A). Among the most represented metabolic categories across all phages, amino acid metabolism and the metabolism of cofactors and vitamins contained 121 and 88 AMGs, respectively, with energy metabolism being the next largest category with 22 AMG hits (Figure S8B). These general pathway results indicate that phages in the intestine presumably affect host metabolism through the consumption of metabolic resources needed for their own biogenesis, as described in phage-host infection studies of model pathogens [58-60] and marine virocells [61].

To further probe all metabolic phage-encoded functions corresponding to our sample cohorts, we assessed all AMG hits for total KEGG pathway abundances. Hierarchical clustering grouped AMGs into 5 distinct metabolic clusters relative to HC and at-risk CCP cohorts (Figure 6A). Among these groups, the gene coding for *phnP* (K06167) stands apart from the others, both in terms of clustering and also for relative pathway abundance (Figure 6A). Among group-associated differences in AMG pathway abundances, there are notable absences among both CCP+ and CCP- individuals. Namely, several clustered transferases such as the mannose-phosphate transferases (*algA, xanB, rfbA, wbpW, pslB*), manno-heptose transferases (*gmhC, hldE, waaE, rfaE*), and the *galE* epimerase and *glmS* transaminase (Figure 6A). Considering the impact of such transferases on bacterial cell wall polysaccharides and biofilm formation [62, 63], these results point to a baseline of phage-driven bacterial surface modifications from HC-derived phages. Conversely, AMGs involved in lipopolysaccharide (LPS) biosynthesis such as the *waaL* O-antigen ligase and the *gmhB* phosphatase are only present in CCP+ phages or at greater abundance in CCP+ phages, respectively, indicating a possible role in immune evasion. Within the CCP- cohort, one of the most abundant AMGs, KEGG orthology entry K23144 encoding for a polyketide sugar transferase important in peptidoglycan biosynthesis is completely absent from the HC cohort and present at lower levels for CCP+ samples. Thus, phage-encoded bacterial surface modifying enzymes such as the sugar transferases and LPS/peptidoglycan biosynthetic genes, are differentially represented across the cohorts in this study, which has implications for bacterial fitness in the intestinal ecosystems and their interactions with the immune system.

We next incorporated the AMG characterization of genomes within our curated set of phages to those that were significantly over- or under-abundant in previous differential abundance analyses (Figures 5G, 5H, and 5I). Among the 20 differentially abundant contigs from the CCP+ vs HC pairwise comparison that contained CRISPR spacer-predicted hosts, 8 of these encoded at least one AMG (Figure 6B). The 9 under-abundant phages in this comparison encode 5 AMGs, including manno-heptose transferases (*gmhC, hldE, waaE, rfaE*), mannose-1-phostphate transferases (*algA, xanB, rfbA, wbpW, pslB*) and *ahbD* AdoMet-dependent heme synthase all together on 1 contig, and *cysH* and *iscS* genes on 2 other contigs (Figure 6B). Among the 11 significantly over-abundant contigs, 3 of these encode the *phnP* phosphodiesterase; 3 phages predicted to infect *Flavonifractor* sp. (Ruminococcaceae) and one predicted to infect Clostridiales bacteria. The remaining AMG found in CCP+-associated over-abundant phages encodes for a cobalamin biosynthesis protein *cobS*, found in marine cyanophages [64], viruses of marine archaea [65], and is considered a core component of marine phage genomes [66], but also ubiquitous in phage genomes that infect *E. coli* [67]. Our identification of a CCP+ over-abundant phage contig targeting *Bacteroides fragilis* and carrying the *cobS* AMG (Figure 6B) reinforces the universal nature of this central AMG that is conserved across hosts and environments [37].

We also identified 16 unique phage contigs with definitive CRISPR spacer-predicted hosts that were differentially abundant and associated with the CCP- cohort (Figure 6C). Within these contigs, 9 are significantly under-abundant compared to healthy controls, with 3 of these encoding AMGs. CCP- associated phages were identified as carrying *cobS*, *DNMT3A*, *thiF*, and *iscS* metabolic genes (Figure 6C). Thus, in contrast to CCP+ associated contigs which harbored *phnP* and *cobS* on a combination of Lachnospiraceae, Rumminococcaceae, and

Bacteroidaceae targeting phages, CCP- associated phages were identified to target primarily Bacteroidaceae and *Actinomyces oris* and harbor a combination of AMGs.

DISCUSSION

RA is a complex disease with an unknown etiology that puts a burden on quality of life resulting in a strong societal impact [68, 69]. In addition to multiple epidemiological factors being associated with RA development, including genetic and familial risk, environmental risk factors and biological sex [3], the microbiota remains an important and understudied factor that likely influences RA autoimmunity [70]. Given the widespread occurrence and diversity of phages in the human intestinal microbiota and their impact on intestinal microbial ecology during health and disease [19, 20, 71], we analyzed this previously neglected component of the microbiota as it relates to RA etiopathogenesis. We used shotgun metagenomics to identify intestinal phages of individuals at risk for developing RA and discovered an association of distinct phage communities with RA-specific serology in the at-risk population.

Using three separate database-independent approaches, we describe a collection of 660 phage genomic sequences, their potential metabolic capability, and their differential abundance. Through a combination of CRISPR spacer matching and Markov clustering with other viral metagenomic sequences from diverse environments, we predicted host assignments for 285 or 43.2% of these phages, which is a high level of taxonomic assignments relative to recent reports of approximately 10 – 30% host assignment identification [23, 48, 72]. By analyzing a core set of *de novo* assembled phage contigs paired with taxonomy, we identified differential phage communities associated with the at-risk RA individuals compared to healthy controls, all while adding novel phage-host assignments to previously unidentified intestinal phages [73, 74].

Phage-host assignments were dominated by Lachnospiraceae-targeting phages, some of which were over-abundant in CCP+ individuals. This expansion of phages also correlated with increased abundances of Lachnospiraceae bacteria in the CCP+ cohort compared to either

CCP- or the healthy cohort, suggesting a link to this family of Firmicutes and CCP autoantibody production in the human intestine. Interestingly, increased abundance of Lachnospiraceae has been observed in at least two previous studies of intestinal microbiotas in mice during the course of collagen-induced arthritis (CIA) [75, 76]. Considering the precedence for overlap of identified phage contigs from mouse intestines to human-associated intestinal phages [23], the previously-reported increase in abundance of Lachnospiraceae bacteria during experimental arthritis in mice is supported by our findings of increased Lachnospiraceae phage-host interactions in CCP seropositive individuals at-risk for developing RA. To this end, given that the FDR individuals included in this study do not show clinical signs of established RA, our identification of a preclinical cohort with increased Lachnospiraceae phage-host interactions could serve as a biological indicator of disease. Similarly, an expansion of Bacteroidaceae-targeting phages associated with the CCP- cohort was described, which corresponds to a previously observed expansion of Bacteroidaceae bacteria following CIA in mice [75]. In addition to these phages serving as potential biomarkers of disease in humans at risk for RA, our data indicate that Bacteroidaceae and Lachnospiraceae-targeting phages designate a distinction between CCP serology status that may serve as an additional indicator of disease progression and/or future disease severity [77]. Notably, bacteria in the Lachnospiraceae and Ruminococcaceae families have been linked to the pre-diabetic intestinal microbiota and diabetic pathogenesis, while Bacteroidaceae are associated with disease protection in a murine model of diabetes [78]. The identification of cohort-specific phage-host interactions sheds light on potential preclinical biomarkers connecting specific dysbiotic intestinal microbial communities to possible regulation of microbiota-mediated mucosal inflammation [1, 79].

We calculated the differential abundance of curated phages on a contig-to-contig basis to estimate dispersion and fold changes of quantitative read mapping matrices. In doing so, we identified 178 differentially abundant contigs (27% of the total curated list) across three pair-wise cohort comparisons. Among the CCP+ vs HC comparison, we observed over-abundant phages targeting *Clostridium scindens, Flavonifractor* sp.*, Actinomyces oris*, as well as other family-level taxonomic assignments. A member of the Lachnospiraceae, *C. scindens* is an intestinal commensal bacterium involved in maintaining homeostatic large intestinal bile acid composition and providing host protection from opportunistic *Clostridioides difficile* blooms [80, 81]. A differential abundance of phage targeting *C. scindens* in the CCP+ at-risk cohort, may have implications for bile acid dysmetabolism in these individuals, which has consequences for inflammatory bowel diseases [82, 83]. Differential abundance of phages in the CCP- cohort revealed several phages targeting Bacteroidaceae and *Bacteroides* species, bacteria involved in multiple reactions of bile acid metabolism promoting host metabolic health [84, 85]. Recent phage-*Bacteroides* interactions have described the influence of phage BV01 in reducing *Bacteroides* bile acid metabolism [86], which has implications for the impact of phages on mammalian gut metabolic function. Our findings suggest individuals at risk for RA harbor divergent communities of phages with potential to alter intestinal metabolic potential through either reduction of key bacterial species and thus reducing endogenous metabolic function, or through the phage-derived introduction of specific AMGs.

Changes to the intestinal metabolome can lead to compositional microbiota transitions that in turn impact host nutrient uptake and immune homeostasis [87]. Considering that manipulations of microbial metabolic pathways in the intestine can influence inflammation and dysbiosis [88], our identification of phage communities with differential abundances of encoded AMGs points to divergent metabolic landscapes associated with at-risk RA cohorts. A majority of the AMGs identified in our analysis make up a group of 14 genes conserved across many environments [37], indicating their functional importance in core metabolism. We were surprised to identify three phages that were over-abundant in the CCP+ cohort (3 of 11 in total), three with *Flavonifractor sp*. predicted hosts and one Clostridiales-targeting phage, encoding the *phnP* phosphodiesterase. Encoding a phosphoribosyl 1,2-cyclic phosphate phosphodiesterase, *phnP* accounts for 10% of the total AMGs represented in our phage genomes, and is differentially abundant among the CCP+ cohort samples. While *phnP* is one of 14 genes considered to be globally conserved across multiple environments [37], it is the only gene among AMGs in our analysis that is the lone representative of its pathway. The PhnP phosphodiesterase, part of a 14-gene operon originally described in *Escherichia coli,* plays a crucial intermediary role in the carbon-phosphorous lyase pathway by degrading a dead-end cyclic phosphonate byproduct [89]. The uniform presence of *phnP* across phages derived from at-risk and healthy cohorts (Figure 6A), suggests phage-driven organophosphonate degradation, which is fundamental for bacteria in diverse environments [90].

Phosphonate degradation is important for phosphorus assimilation in enteric bacteria [91], although phosphonate metabolism has not been described for *Flavonifractor* species and a *phnP* homolog is not available for this genus in the KEGG database (K06167). In a recent study characterizing microbiota KEGG orthologs as predictors of methotrexate responsiveness for RA treatment, a gene in the phosphonate transport system, *phnC* (K02041), exhibited high median random forest importance as a predictor of drug response in new-onset RA subjects [92]. The contribution of the phosponate metabolic pathway in bacteria and phages, will require further exploration in the context of RA pathogenesis and treatment. However, it is possible that these phage-encoded metabolic products are supplementing phosphorous uptake among Ruminococcaceae and Lachnospiraceae bacteria that predominate in CCP+ individuals prior to RA clinical symptoms. Our analysis is limited in that we did not measure a longitudinal progression of microbial metabolic pathways in these human samples, yet these metabolic associations warrant further investigations into causality and the potentially cascading effects on interbacterial interactions [93].

Our results point to divergent communities of phages with multiple bacterial host targets that group according to anti-CCP serology in individuals predisposed to developing RA. These at-risk individuals who develop seropositive RA, a disease manifestation that is more severe [94] and less responsive to treatment [95], endure a prolonged asymptomatic period before pathological early RA develops in those who are at a higher disease susceptibility in the preclinical RA state [1]. Current approaches for RA diagnosis rely in large part on anti-CCP serology which has up to 93% specificity but as low as 67% sensitivity (for the CCP3.1 assay used here) [39], indicating that a negative result does not preclude current or development of clinically apparent RA. Phage community composition analyses may complement existing diagnoses for RA, considering that intestinal phages can play important roles in immune tolerance, mucosal immunity, and microbial homeostasis [96]. Given that phage community alterations have been shown to precede autoimmunity development in children at risk for developing type 1 diabetes [26], phage community structure should be considered as a biomarker for diseases such as RA that are influenced by non-genetic microbial factors [19]. To that end, we have characterized the intestinal viromes of RA at-risk individuals corresponding to anti-CCP serology status. Furthermore, we calculated species-specific phage-host interactions and identified over-abundances of *C. scindens* and *A. oris* targeting phages in CCP+ and CCP- individuals, respectively. Divergent metabolic profiles evident by differential abundance of AMG-encoding phages in both conditions warrant further interrogation during models of RA-like disease. Future work should investigate the potential of phages in a murine CIA model to determine the influence of RA-associated phages with immunomodulation and inflammatory disease progression. Our multifaced approaches for phage prediction and phage host assignments hold promise to better ascertain the occurrence and diversity of the virome and the identification of key phages influencing the microbiota and individuals at risk for developing RA autoimmune disease. This RA-focused study implicating specific phage populations could open new avenues to assess the basis for phage implication in other microbiota dysbiosis-associated diseases.

RESOURCE AVAILABILITY

Lead Contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Breck A. Duerkop (breck.duerkop@cuanschutz.edu).

Materials Availability

This study did not generate new unique reagents.

Data and Code Availability

The VLP and whole metagenome DNA sequencing reads as well as the final curated phage contigs generated in this study are available at the European Nucleotide Archive under the Study titled “Intestinal VLP reads and predicted phage contigs for at-risk RA individuals” (accession numbers PRJEB42612 and ERP126498). The VLP and whole metagenome raw unmapped read sets are available for each of the 25 individual samples included in this study and are available under the Study Primary Accession PRJEB42612. The 660 curated contigs are compiled in a multifasta file deposited as Sample SAMEA7856466 under the same Study PRJEB42612.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Study Subjects and Fecal Samples

Fecal samples were obtained from individuals recruited for the SERA (Studies of the Etiology of Rheumatoid Arthritis) initiative, aimed at understanding the mechanisms that prelude the preclinical development of RA. SERA is a multicenter prospective cohort study that has identified first-degree relative (FDR) probands defined as a parent, full sibling, or offspring of individuals with diagnosed clinical RA [38]. FDR probands were evaluated in extensive clinical research visits, longitudinal follow-ups, and autoantibody testing to determine CCP status [38]. FDR probands were split into cohorts dependent on serum CCP levels, with 100% of subjects in the CCP+ cohort positive and 0% of subjects in either CCP- or HC (Healthy Control) cohorts testing positive. Healthy control subjects were recruited and included in this study as described previously [97]. The present study consisted of 25 subjects split into 3 cohorts, of which 8 were CCP+, 8 were CCP-, and 9 were HC. Ethical approval for this study was obtained from the University of Colorado Multiple Institutional Review Board (COMIRB) study numbers 01-675 (primary) and also 13-2606 and 14-1751. COMIRB Protocol 01-675 included informed consent with HIPAA authorization for stool sample collections. Stool samples were obtained independently by SERA study participants and returned within 1 week of their original visit. Samples were stored in aliquots at -20°C until processing.

METHODS

Extraction of Fecal Whole Metagenome and VLP DNA, Library Preparation and Sequencing

Whole metagenome and VLP fraction DNA were isolated as described previously [98], with some modifications as follows. For all samples, 0.1 g of human stool was homogenized in 8 mL salt magnesium plus (SM+) buffer [99] and 0.5 ml of homogenate was transferred to a BashingBead Lysis tube (Zymo) and designated as the whole metagenome sample. Whole metagenome DNA was extracted using a ZymoBIOMICS DNA kit (Zymo) following the manufacturer recommended protocol. VLPs were clarified from the remaining 7.5 mL of sample by three successive centrifugation steps (3200g for 10 min, 3200g for 10 min, 7800g for 10 min), and the supernatant was filtered through a 0.45-µm PVDF filter membrane. VLPs were precipitated by adding 0.5M NaCl and 10% wt/vol PEG8000 and incubating on ice at 4ºC overnight, followed by centrifugation (7800g for 20 min). VLP pellets were resuspended in 400 µL SM+ buffer and treated with 40 µL DNase buffer (10 mM $CaCl_2$, 50 mM $MgCl_2$), 25 units DNase, and 15 units RNase for 1 hr at 37ºC. VLPs were further treated with 50 mg/mL proteinase K and 0.5% SDS for 30 min at 56ºC before addition of 100 µL phage lysis buffer (4.5 M guanidiniumisothiocyanate, 44 mM sodium citrate pH 7.0, 0.88% sarkosyl, 0.72% 2-

mercaptoethanol) and incubated for 10 min at 65ºC. VLP DNA was precipitated and extracted with an equal volume of phenol/chloroform/isoamyl alcohol 25:24:1, spun at 7800g for 5 min, and further extracted with an equal volume of chloroform. VLP DNA was precipitated with 0.3M NaOAc (pH 5.2) and an equal volume of isopropanol, washed with ice-cold 70% ethanol, and resuspended in sterile water.

Metagenomic DNA Sequencing

VLP and whole metagenomic DNA was sequenced using the Illumina NovaSEQ 6000 platform with paired-end 150-cycle sequencing chemistry. DNA libraries were amplified using the Ovation Ultralow System v2 (Nugen, part no. 0334) library preparation kit including 12 cycles of amplification. TruSeq adapters (Illumina) were used for multiplexing. Libraries were quantified using a Qubit and quality was measured using a Tapestation. All library preparation, quantification, quality assessment and control, were performed by the University of Colorado Anschutz Medical Campus Genomics and Microarray Core.

16S rRNA Amplicon Sequencing and Analysis

16S rRNA gene analysis was performed using fecal samples that were processed for isolation of whole metagenomic DNA using a ZymoBIOMICS DNA kit (Zymo) and stored at -80°C. Amplicons of the 16S rRNA gene V4 region were amplified using Earth Microbiome Project primers 515F and 806R [100] with custom barcodes. Samples were sequenced on the Illumina MiSeq platform with paired end 250 bp reads using bTEFAP technology [101] by MR DNA (Molecular Research LP, Shallowater, TX), and processed using mothur v.1.44.0 [102]. Sequenced reads, which averaged 607,915 ± 112,641.7 per sample, were demultiplexed, assembled as contigs, and processed to remove chimeras and erroneous sequences per the Kozich protocol [103] and mothur MiSeq SOP (https://mothur.org/wiki/miseq_sop/ accessed

07/16/2020). Sequences were aligned to the Greengenes core reference alignment for taxonomy using the 2013 release (gg_13_8_99) [104]. Sequences were differentiated into amplicon sequence variants (ASVs) using the make.shared command, resulting in a total of 8,108,071 sequences. Subsampling was performed using 186,745 sequences, which corresponded to the smallest sample in our dataset. Diversity measurements were performed using mothur calculators to estimate community richness (Chao1 estimator), community evenness (Shannon evenness), and community diversity (inverse Simpson index).

Decontamination and Read Processing

Metagenomic reads were decontaminated and trimmed as previously described [23] using BBMap short read aligner v38.56 [105]. Briefly, raw reads were mapped to the internal Illumina phage genome control phiX174 (J02482.1), human reference genome (hg38), and potential laboratory contaminants including mouse genome (mm10), *Enterococcus faecalis* V583 genome (NC_004668.1), *E. faecalis* OG1RF genome (NC_017316.1), and *E. faecalis* phage VPE25 (LT546030.1) using the bbsplit algorithm with default settings. Unmapped reads were trimmed of adapter sequences, with low quality reads and reads of insufficient length removed using the bbduk algorithm with the following parameters: ktrim = lr, k = 20, mink = 4, minlength = 20, qtrim = f, as previously described [23].

Metagenomic Assembly

Decontaminated and trimmed R1 and R2 reads were interleaved using the fq2fa --merge command from the IDBA-UD package [106]. Whole metagenome and VLP assemblies were performed using the MEGAHIT assembler v1.2.7 [107] using the default setting plus the following flags: --presets meta-large (--k-min 27 --k-max 127 --k-step 10) for large and complex metagenomes.

Quantitative Read Mapping and Construction of the Curated VLP Contig Database

VLP reads were assembled into 25 individual sample sets, corresponding to the 25 individual fecal samples included in our study. All contigs resulting from MEGAHIT assembly were filtered to a minimum length of 5kb, resulting in a pool of 80,762 total contigs from all samples. Three separate independently published methods were employed to identify putative phages from the pooled set of contigs over 5kb in length. First, the P/M read mapping approach was used whereby each sample's VLP and whole metagenome reads were mapped to their corresponding assembled contigs, using BBMap as previously described [23]. After pooling, the top 100 largest ratios of VLP reads to whole metagenome reads for all 25 read-mapping sets for each sample were identified and pooled. Redundancy was removed using cd-hit-est at an identity threshold of 95% resulting in 2117 unique contig sequences. Next, as a separate method, putative phages from the 80,762 contigs were identified by searching for viral protein family (VPF) hits, as previously described [41]. Separate filters were applied for VPF hits calculated in relation to total genes, microbial genes, and percent non-viral Pfams. 2,902 contigs were identified that contained 5 or more VPF hits and with non-viral Pfam hits below 20%. 263 contigs were identified with 5 or more VPF hits, with more viral gene content than microbial genes per contig, and 644 contigs were identified with 2 – 4 VPF hits and 0 microbial gene hits. Finally, 976 contigs were identified with only 1 VPF hit per contig, and were included regardless of microbial gene content. The third and final independent phage contig identification method used was VIBRANT v1.2.1 [37], a neural network machine learning algorithm that identifies viral protein signatures. VIBRANT identified 4,758 unique phage contigs. After combining these three independent approaches used to identify unique sets of phages, all sets were combined and the overlapping 660 contigs were used for analysis as the curated contig set. To assess contig completion and contamination levels, CheckV v0.6.0 was used with standard operating parameters.

Differential Abundance Analyses

To calculate differential abundance in pairwise analyses, we first generated read mapping count matrices by mapping all VLP reads to the curated contig set of 660 contigs. The bbmap algorithm from the BBMap suite of tools was used with the following parameters: ambiguous = random, qtrim = lr, minid = 0.97. Total raw read counts were aggregated per contig and assembled into 25 count matrices for all samples, which were then used as input for DESeq2 v1.20.0 [53] running in R version 3.6.3 for comprehensive differential abundance analysis. Raw un-normalized read count coverage values were used to compare fold changes across three pairwise comparisons: CCP+ vs. HC, CCP- vs. HC, and CCP+ vs. CCP- groups. The standard workflow for differential analysis within DESeq2 was used, producing logarithmic fold-change values incorporating Wald tests for *p*-value calculations and the Benjamini-Hochberg multiple testing correction for the adjusted *p*-value. In total, 178 phage contigs from our set of 660 were found to be differentially abundant using thresholds of $\log_2$ Fold Change < -1 or > 1 and adjusted p-value < 0.001.

VLP Clustering, Phage Host Matching, and AMG Identification

Clustering of all viral contigs within the RA virome described in this study was performed using two lists of contigs, the total 4,785 viral sequences identified by all filters of the VPF method, as well as the final curated set of 660 contigs. First, all 4,785 sequences were screened against the most recent iteration of the public viral database IMG/VR v3.0 [44] using blastn with 95% sequence similarity over 85% of each 1kb region of the contig, which resulted in 19,892 viral sequences. Then, a total of 24,926 sequences were screened against each other using blastn with the same parameters and omitting duplicate hits. Markov clustering of these 9.4 million connections resulted in a total of 1,193 clusters encompassing 22,306 total sequences. Overall, 2,420 of the 4,785 total RA virome sequences were clustered into 1,184 clusters. Of these clusters, 41 contained a reference viral isolate, 1,037 contained another metagenomic viral contig from IMG/VR, and 106 were identified as originating from RA metagenomic sequencing projects. Lastly, clustering was also calculated for the 660 curated viral sequences, which resulted in 266 individual clusters containing 336 (roughly 48% of curated set) unique sequences. Phage host assignments were determined via bacterial CRISPR spacer matching as previously described [23], requiring at least 93% sequence identity match over the entire spacer length and allowing for up to 2 mismatches. Of our 660 curated contig list, 207 (31.4%) had CRISPR spacers matching reference isolates therefore leading to host predictions for a third of our final contigs. VIBRANT v1.2.1 was used to identify auxiliary metabolic genes (AMGs) according to KEGG metabolic pathway annotations. VIBRANT annotates using VOG, Pfam, and KEGG databases; therefore, if the best HMM hit is to the KEGG database and also if the annotation is in a metabolic pathway, the hit gets called as an AMG.

Data Visualizations

Various R packages were used, including DESEq2, ggplot2, ComplexHeatmap, pheatmap, corrplot, RColorBrewer, and EnhancedVolcano. Graphpad Prism v8.4.3 was used for all supplemental calculations. Lastly, SankeyMATIC (https://github.com/nowthis/sankeymatic) and meta-chart (https://www.meta-chart.com/venn) were used to create the Sankey and Venn diagrams depicted in Figure 1, respectively.

FIGURE LEGENDS

Figure 1. Generation and curation of *de novo* assembled VLP contigs. Metagenomic sequencing was carried out for 25 samples belonging to 3 cohorts of individuals, FDRs at risk for developing RA later in life with either CCP+ or CCP- serology status, and a Healthy Control (HC) group. (A) Contigs were assembled *de novo* for all samples, ranging from 30,011 to 284,689 contigs per sample, and a total of 3,557,500 contigs for the entire sample set. Each node on the Sankey diagram is scaled to the number of contigs it contains. Thresholds of

minimum contig sizes being greater than 1 and 5 kilobases reduced the total numbers to 564,228 and 80,762 contigs respective to the size cut-off. Three independent methods were used to identify putative phages from the list of 80,762 contigs (boxed portion of panel A), resulting in 2,117 contigs from the P/M ratio method, 4,785 contigs from the Viral Protein Families method, and 4,758 contigs using the VIBRANT algorithm. (B) A Venn diagram was created to show the overlap of redundant contigs identified among the three methods. 660 unique contigs were identified independently by all phage identification methods. (C) CheckV analysis of the three separate methods as well as the final set of curated contigs revealed a disparity in host contamination, with the set of 660 contigs being relatively free of host bacterial contamination. Colors were assigned to the CheckV categories that account for prophage elements based on their position on the contig sequence, as well as pure viral (green) and pure bacterial (grey) classifications.

Figure 2. Clustering with metagenomic viral contigs reveals viral ecological composition. (A) Host assignments for the set of curated phages based on Markov clustering to the IMG/VR database metagenomic viral clusters or direct match to bacterial CRISPR spacers, based on cohort abundance. Bacteroidaceae, Lachnospiraceae, Ruminococcaceae, and Streptococcaceae hosts are evident to be cumulatively more abundant than other bacterial families, especially for the CCP+ cohort. (B) Cladogram of the complete host phylogeny at the genus level for all spacers identified from total RA virome via the VPF method. The pie chart at the center represents all 958 CRISPR spacers from the family level quantified in panel A that have been color coordinated on this cladogram as well. Total host hits were quantified at the genus level and are represented in relative size by colored circles, indicating host assignments that were discerned via clustering (dark grey) and those that were identified via direct CRISPR spacer matching (light grey). Total CRISPR spacers per contig with family level host taxonomy

assignments were tabulated per cohort group (C) and differentiated as narrow or broad phage host ranges (D) based on target uniformity to bacterial family.

Figure 3. Phage-host assignments for curated VLP contigs reveal cohort-based differential read recruitment among several bacterial families. Relative abundances were calculated for all VLP reads mapped to phages predicted to target Bacteroidaceae (A), Clostridiaceae (B), Lachnospiraceae (C), Ruminococcaceae (D), Streptococcaceae (E), and Veillonellaceae (F) bacterial families. For these analyses, VLP reads were mapped to predicted phage contigs to which CRISPR spacers were assigned using bbmap at a 97% minimum read-mapping identity level. Scaffold abundances were averaged across all samples and statistics were determined by nonparametric Wilcoxon tests (* $p < 0.05$, ** $p < 0.01$, **** $p < 0.0001$).

Figure 4. CRISPR spacer host metadata reveal CCP+ phages represent greater variability in microbial host ecology. Phage host isolate ecology metadata was compiled from JGI/GOLD v7.0 and broken down by Ecosystem, Ecosystem Category, Ecosystem Type, and Ecosystem Subtype distributions accordingly for all CRISPR spacers identified within our list of 660 phages. (A) Ecosystem Distribution showing the percent host-associated spacers calculated for each contig based on cohort distribution. (B) Ecosystem Category distribution showing the percent human-associated spacers. (C) Ecosystem Type distribution showing the percent of contigs that contain spacers originating from the digestive system. (D) Ecosystem Subtype showing the percent of contigs that contain spacers originating from the large intestine microenvironment. Cohort distributions based on these metadata revealed a disproportionate distribution of CRISPR spacers among samples originating from CCP+ individuals when compared to CCP- or HC groups. Statistical significance was determined using pairwise Wilcoxon rank sum tests for comparisons between the three groups, using the Benjamini-Hochberg correction for multiple testing comparisons (* $p = 0.023$, **** $p < 2 \times 10^{-16}$).

Figure 5. Quantitative read mapping exposes differentially abundant contigs despite sample cohesiveness. Quantitative read mapping of all VLP read sets to the final curated 660 phages reveals contig-contig dissimilarities despite minimal sample-sample variance or intra-sample hierarchical clustering. Differential abundance calculations were carried out within the DESeq2 package by way of 3 pairwise comparisons: CCP+ vs. HC, CCP- vs. HC, and CCP+ vs. CCP-. (A, B, C) Analyses of the first and second principal components for sample-to-sample exploratory analyses revealed minimal variance explained across all comparisons. (D, E, F) Euclidian distances for sample-sample read-based coverages were used for hierarchical clustering across all pairwise comparisons reveal minimal clustering based on sample type. (G, H, I) Volcano plots reveal 9%, 10%, and 8% of contigs included in our analysis are differentially abundant respective to CCP+ vs. HC, CCP- vs. HC, and CCP+ vs. CCP- group-based comparisons of specific contig community members.

Figure 6. Phage auxiliary metabolic gene abundances highlight cohort-associated disparities in potential metabolic function. AMGs were identified within the VIBRANT algorithm, based on screening 2,835 auxiliary metabolic genes with KEGG Orthology annotations (March 2019 KEGG release, ftp://ftp.genome.jp/pub/db/kofam/archives/2019-03-20/). (A) Total counts per KEGG Pathway were used normalize relative abundance of AMGs per sample, which were clustered using the ComplexHeatmap package in R. Areas in black indicate no AMG hits were present for the entire cohort for the 660 contig samples. (B) Differentially abundant contig for the CCP+ to HC pairwise comparison, visualizing only the contigs which had CRISPR spacer-predicted hosts. Color-coded stars belong to a list of AMGs and indicate association with the contig they are adjacent to. (C) Differentially abundant contigs for the CCP- vs HC comparison.

TABLES

Table 1. Summary of the Subjects' Characteristics for the Samples Included in the Study

VARIABLE	HC	CCP+	CCP-
Count	9	8	8
Age (mean)	44.4	61.3	49
Age (SD)	13.6	11	15.7
Sex (% female)	66.7	88.9	62.5
Serum CCP+ (%)	0	100	0
Ever smokers (%)	22.2	33.3	0

SUPPLEMENTAL INFORMATION

Figure S1. Overview of methods for VLP isolation and phage identification from sequencing reads. (A) Individual stool samples were homogenized and split into P and M subsamples for generating VLP and whole metagenome DNA, respectively. (B) Total sequencing reads generated per sample for each P and M read sets, after quality control and decontamination. (C) Total assembled contigs with length greater than 5kb generated per sample for each P and M read sets. (D) Overview of the computational pipeline used to identify phages; from short-insert pair end read sets averaging approximately 75M read pairs per sample, to the 80,762 *de novo* assembled contigs greater than 5kb in length, and the three independent methods for phage identification (P/M, VPF, VIBRANT).

Figure S2. Estimation of contig completeness by CheckV. Distribution of contig lengths across contig quality categories according to the MIUViG standards. Contigs derived from the (A) P/M ratio method of phage identification, (B) the VPF method, (C) VIBRANT algorithm, and finally (D) the curated contig list. Boxplots depict the following five summary statistics: median, lower and upper hinges corresponding to the first and third quartiles, and two whiskers corresponding to 1.5 times the interquartile range between the first and third quartiles. Points beyond the whiskers correspond to outlier points.

Figure S3. Lifestyle and morphology distributions of curated phage contigs. (A) Total contigs per sample among the three groups, divided according to infection mechanism (lytic vs. lysogenic) as determined by the VIBRANT algorithm. (B) Relative abundances of phage lifestyles as determined by the VIBRANT algorithm. In total, for our 660 predicted phages, 467 (70.8%) are classified as lytic and 193 (29.2%) are classified as lysogenic by VIBRANT. (C) Viral taxonomy of all contigs per sample including the top four morphotypes: Siphoviridae, Myoviridae, Podoviridae, and Microviridae. (D) Relative abundance of all viral morphotypes identified for all 660 phages. Viral taxonomy was determined using a custom database described in this preprint by Kieft et al. (2020; bioRxiv preprint doi: https://doi.org/10.1101/2020.08.24.253096).

Figure S4. Clustering distribution of singletons and viral groups. (A) Total singletons, viral contigs that did not cluster with any other genome, per sample and RA cohort group. (B) Distribution of total viral clusters in relation to the number of viral genomes clustered within each group.

Figure S5. CRISPR spacer host metadata distribution of environmental and engineered derived phages per cohort. Phage host isolate ecology metadata was compiled from JGI/GOLD v7.0 at the highest Ecosystem classification level for all CRISPR spacers identified within our list of 660 phages. Data is presented as percent of spacers per contig whose metadata is designated as originating from (A) environmental or (B) engineered environments, distributed across the three RA cohort groups. Statistical significance was determined using pairwise Wilcoxon rank sum tests for comparisons between the three groups, using the Benjamini-Hochberg correction for multiple testing comparisons (* $p = 0.011$, **** $p < 2 \times 10^{-16}$).

Figure S6. Principal component analyses based on quality, predicted phage lifestyle, and sample cohort. Principal components for the final curated set of 660 contigs derived from the VIBRANT phage identification program categorized by (A) contig quality, (B) phage lifestyle, and (C) cohort group. Total identified open reading frames were incorporated in analyses in (A) and (B), showing a greater dispersion of smaller sized contigs and a consensus grouping of bigger contigs.

Figure S7. Analysis of bacterial family diversity from fecal samples based on 16S sequencing and analyzed using mothur. (A) Relative abundances of bacterial families based on ASV binning reveals a significant difference in Lachnospiraceae bacteria originating from CCP+ fecal DNA samples. Unpaired nonparametric Mann-Whitney tests were used to compare ranks, revealing *p* values of 0.0464 comparing CCP+ to HC individuals. Community richness was measured by the standard observed richness calculator in mothur (B) as well as the Chao1 richness estimate (C). Community evenness was measured using the Shannon index (D), and community diversity was measured using the inverse Simpson index (E). No statistically significant differences were observed among any of the above calculators using nonparametric tests of significance among the three groups.

Figure S8. Distribution of auxiliary metabolic genes found on curated contigs. (A) A total of 252 AMGs were discovered among our 660 phages, distributed across the three cohorts. (B) AMGs were categorized predominantly as belonging to amino acid and cofactor/vitamin metabolism categories.

ACKNOWLEDGEMENTS

The authors thank all participants in the SERA studies at the University of Colorado and elsewhere. We thank Rodolphe Barrangou for comments and suggestions to improve this manuscript. We thank Monica Ransom and Katrina Diener from the University of Colorado Anschutz Medical Campus Genomics and Microarray Core for performing Illumina library preparation and sequencing. The following funding sources supported this work: NIH U01AI101981 (V.M.H., K.A.K.), NIH U01AI130830 (B.A.D.), startup funds from the University of Colorado School of Medicine (B.A.D.), NIH Research Training in Rheumatology T32AR007534 (M.R.M.), Rheumatology Research Foundation Future Scientist Award (M.E.C.).

AUTHOR CONTRIBUTIONS

Conceptualization, M.R.M, D.P., A.C., K.A.K., and B.A.D.; Methodology, M.R.M., D.P., K.K., A.C., J.A.S., M.L.F., M.K.D., and B.A.D.; Investigation, M.R.M, D.P., K.K., A.C., M.E.C.; Sample Procurement, M.E.C., J.A.S., M.L.F., and M.K.D.; Visualization, M.R.M, D.P. and K.K.; Writing – Original Draft, M.R.M and B.A.D.; Writing – Review & Editing, M.R.M, D.P., K.K., A.C., M.E.C, A.S., K.D.D., V.M.H., K.A.K. and B.A.D.; Funding Acquisition, V.M.H. and B.A.D.; Resources, B.A.D.; Supervision, B.A.D., V.M.H., K.D.D., K.A., A.S. and K.A.K.

DECLARATION OF INTERESTS

D.P.E is an employee of Mammoth Biosciences and co-founder/employee of Ancilia Therapeutics. A.S. is the founder/employee of Ancilia Therapeutics. B.A.D. is a co-founder and shareholder of Ancilia Therapeutics.

REFERENCES

1. Holers VM, Demoruelle MK, Kuhn KA, Buckner JH, Robinson WH, Okamoto Y, Norris JM, Deane KD. Rheumatoid arthritis and the mucosal origins hypothesis: protection turns to destruction. Nat Rev Rheumatol. 2018;14(9):542-57.

2. MacGregor AJ, Snieder H, Rigby AS, Koskenvuo M, Kaprio J, Aho K, Silman AJ. Characterizing the quantitative genetic contribution to rheumatoid arthritis using data from twins. Arthritis Rheum. 2000;43(1):30-7.

3. Deane KD, Demoruelle MK, Kelmenson LB, Kuhn KA, Norris JM, Holers VM. Genetic and environmental risk factors for rheumatoid arthritis. Best Pract Res Clin Rheumatol. 2017;31(1):3-18.

4. Chang HH, Liu GY, Dwivedi N, Sun B, Okamoto Y, Kinslow JD, Deane KD, Demoruelle MK, Norris JM, Thompson PR, Sparks JA, Rao DA, Karlson EW, Hung HC, Holers VM, Ho IC. A molecular signature of preclinical rheumatoid arthritis triggered by dysregulated PTPN22. JCI Insight. 2016;1(17):e90045.

5. Demoruelle MK. Mucosa biology and the development of rheumatoid arthritis: potential for prevention by targeting mucosal processes. Clin Ther. 2019;41(7):1270-8.

6. Demoruelle MK, Bowers E, Lahey LJ, Sokolove J, Purmalek M, Seto NL, Weisman MH, Norris JM, Kaplan MJ, Holers VM, Robinson WH, Deane KD. Antibody responses to citrullinated and noncitrullinated antigens in the sputum of subjects with rheumatoid arthritis and subjects at risk for development of rheumatoid arthritis. Arthritis Rheumatol. 2018;70(4):516-27.

7. Demoruelle MK, Harrall KK, Ho L, Purmalek MM, Seto NL, Rothfuss HM, Weisman MH, Solomon JJ, Fischer A, Okamoto Y, Kelmenson LB, Parish MC, Feser M, Fleischer C, Anderson C, Mahler M, Norris JM, Kaplan MJ, Cherrington BD, Holers VM, Deane KD. Anti-citrullinated protein antibodies are associated with neutrophil extracellular traps in the sputum in relatives of rheumatoid arthritis patients. Arthritis Rheumatol. 2017;69(6):1165-75.

8. Hughes-Austin JM, Deane KD, Derber LA, Kolfenbach JR, Zerbe GO, Sokolove J, Lahey LJ, Weisman MH, Buckner JH, Mikuls TR, O'Dell JR, Keating RM, Gregersen PK, Robinson

WH, Holers VM, Norris JM. Multiple cytokines and chemokines are associated with rheumatoid arthritis-related autoimmunity in first-degree relatives without rheumatoid arthritis: Studies of the Aetiology of Rheumatoid Arthritis (SERA). Ann Rheum Dis. 2013;72(6):901-7.

9. Willis VC, Demoruelle MK, Derber LA, Chartier-Logan CJ, Parish MC, Pedraza IF, Weisman MH, Norris JM, Holers VM, Deane KD. Sputum autoantibodies in patients with established rheumatoid arthritis and subjects at risk of future clinically apparent disease. Arthritis Rheum. 2013;65(10):2545-54.

10. Brusca SB, Abramson SB, Scher JU. Microbiome and mucosal inflammation as extra-articular triggers for rheumatoid arthritis and autoimmunity. Curr Opin Rheumatol. 2014;26(1):101-7.

11. Scher JU, Sczesnak A, Longman RS, Segata N, Ubeda C, Bielski C, Rostron T, Cerundolo V, Pamer EG, Abramson SB, Huttenhower C, Littman DR. Expansion of intestinal *Prevotella copri* correlates with enhanced susceptibility to arthritis. Elife. 2013;2:e01202.

12. Maeda Y, Kurakawa T, Umemoto E, Motooka D, Ito Y, Gotoh K, Hirota K, Matsushita M, Furuta Y, Narazaki M, Sakaguchi N, Kayama H, Nakamura S, Iida T, Saeki Y, Kumanogoh A, Sakaguchi S, Takeda K. Dysbiosis contributes to arthritis development via activation of autoreactive T cells in the intestine. Arthritis Rheumatol. 2016;68(11):2646-61.

13. Zhang X, Zhang D, Jia H, Feng Q, Wang D, Liang D, Wu X, Li J, Tang L, Li Y, Lan Z, Chen B, Li Y, Zhong H, Xie H, Jie Z, Chen W, Tang S, Xu X, Wang X, Cai X, Liu S, Xia Y, Li J, Qiao X, Al-Aama JY, Chen H, Wang L, Wu QJ, Zhang F, Zheng W, Li Y, Zhang M, Luo G, Xue W, Xiao L, Li J, Chen W, Xu X, Yin Y, Yang H, Wang J, Kristiansen K, Liu L, Li T, Huang Q, Li Y, Wang J. The oral and gut microbiomes are perturbed in rheumatoid arthritis and partly normalized after treatment. Nat Med. 2015;21(8):895-905.

14. Vaahtovuo J, Munukka E, Korkeamaki M, Luukkainen R, Toivanen P. Fecal microbiota in early rheumatoid arthritis. J Rheumatol. 2008;35(8):1500-5.

15. Toivanen P, Vartiainen S, Jalava J, Luukkainen R, Mottonen T, Eerola E, Manninen R. Intestinal anaerobic bacteria in early rheumatoid arthritis (RA). Arthritis Research & Therapy. 2002;4.

16. Alpizar-Rodriguez D, Lesker TR, Gronow A, Gilbert B, Raemy E, Lamacchia C, Gabay C, Finckh A, Strowig T. *Prevotella copri* in individuals at risk for rheumatoid arthritis. Ann Rheum Dis. 2019;78(5):590-3.

17. Pianta A, Arvikar S, Strle K, Drouin EE, Wang Q, Costello CE, Steere AC. Evidence of the immune relevance of *Prevotella copri*, a gut microbe, in patients with rheumatoid arthritis. Arthritis Rheumatol. 2017;69(5):964-75.

18. Drago L. *Prevotella copri* and microbiota in rheumatoid arthritis: fully convincing evidence? J Clin Med. 2019;8(11).

19. Duerkop BA. Bacteriophages shift the focus of the mammalian microbiota. PLoS Pathog. 2018;14(10):e1007310.

20. Minot S, Sinha R, Chen J, Li H, Keilbaugh SA, Wu GD, Lewis JD, Bushman FD. The human gut virome: inter-individual variation and dynamic response to diet. Genome Res. 2011;21(10):1616-25.

21. Manrique P, Bolduc B, Walk ST, van der Oost J, de Vos WM, Young MJ. Healthy human gut phageome. Proc Natl Acad Sci U S A. 2016;113(37):10400-5.

22. Shkoporov AN, Hill C. Bacteriophages of the human gut: the “known unknown" of the microbiome. Cell Host Microbe. 2019;25(2):195-209.

23. Duerkop BA, Kleiner M, Paez-Espino D, Zhu W, Bushnell B, Hassell B, Winter SE, Kyrpides NC, Hooper LV. Murine colitis reveals a disease-associated bacteriophage community. Nat Microbiol. 2018;3(9):1023-31.

24. Clooney AG, Sutton TDS, Shkoporov AN, Holohan RK, Daly KM, O'Regan O, Ryan FJ, Draper LA, Plevy SE, Ross RP, Hill C. Whole-virome analysis sheds light on viral dark matter in inflammatory bowel disease. Cell Host Microbe. 2019;26(6):764-78 e5.

25. Khan Mirzaei M, Khan MAA, Ghosh P, Taranu ZE, Taguer M, Ru J, Chowdhury R, Kabir MM, Deng L, Mondal D, Maurice CF. Bacteriophages isolated from stunted children can regulate gut bacterial communities in an age-specific manner. Cell Host Microbe. 2020;27(2):199-212 e5.

26. Zhao G, Vatanen T, Droit L, Park A, Kostic AD, Poon TW, Vlamakis H, Siljander H, Harkonen T, Hamalainen AM, Peet A, Tillmann V, Ilonen J, Wang D, Knip M, Xavier RJ, Virgin HW. Intestinal virome changes precede autoimmunity in type I diabetes-susceptible children. Proc Natl Acad Sci U S A. 2017;114(30):E6166-E75.

27. Barr JJ, Auro R, Furlan M, Whiteson KL, Erb ML, Pogliano J, Stotland A, Wolkowicz R, Cutting AS, Doran KS, Salamon P, Youle M, Rohwer F. Bacteriophage adhering to mucus provide a non-host-derived immunity. Proc Natl Acad Sci U S A. 2013;110(26):10771-6.

28. Quistad SD, Grasis JA, Barr JJ, Rohwer FL. Viruses and the origin of microbiome selection and immunity. ISME J. 2017;11(4):835-40.

29. Gorski A, Dabrowska K, Miedzybrodzki R, Weber-Dabrowska B, Lusiak-Szelachowska M, Jonczyk-Matysiak E, Borysowski J. Phages and immunomodulation. Future Microbiol. 2017;12:905-14.

30. Minot SS, Willis AD. Clustering co-abundant genes identifies components of the gut microbiome that are reproducibly associated with colorectal cancer and inflammatory bowel disease. Microbiome. 2019;7(1):110.

31. Yu AI, Zhao L, Eaton KA, Ho S, Chen J, Poe S, Becker J, Gonzalez A, McKinstry D, Hasso M, Mendoza-Castrejon J, Whitfield J, Koumpouras C, Schloss PD, Martens EC, Chen GY. Gut microbiota modulate CD8 T cell responses to influence colitis-associated tumorigenesis. Cell Rep. 2020;31(1):107471.

32. Norman JM, Handley SA, Baldridge MT, Droit L, Liu CY, Keller BC, Kambal A, Monaco CL, Zhao G, Fleshner P, Stappenbeck TS, McGovern DP, Keshavarzian A, Mutlu EA, Sauk J, Gevers D, Xavier RJ, Wang D, Parkes M, Virgin HW. Disease-specific alterations in the enteric virome in inflammatory bowel disease. Cell. 2015;160(3):447-60.

33. Gevers D, Kugathasan S, Denson LA, Vazquez-Baeza Y, Van Treuren W, Ren B, Schwager E, Knights D, Song SJ, Yassour M, Morgan XC, Kostic AD, Luo C, Gonzalez A, McDonald D, Haberman Y, Walters T, Baker S, Rosh J, Stephens M, Heyman M, Markowitz J, Baldassano R, Griffiths A, Sylvester F, Mack D, Kim S, Crandall W, Hyams J, Huttenhower C, Knight R, Xavier RJ. The treatment-naive microbiome in new-onset Crohn's disease. Cell Host Microbe. 2014;15(3):382-92.

34. Lee JY, Mannaa M, Kim Y, Kim J, Kim GT, Seo YS. Comparative analysis of fecal microbiota composition between rheumatoid arthritis and osteoarthritis patients. Genes (Basel). 2019;10(10).

35. Van Belleghem JD, Dabrowska K, Vaneechoutte M, Barr JJ, Bollyky PL. Interactions between bacteriophage, bacteria, and the mammalian immune system. Viruses. 2018;11(1).

36. Rodriguez-Valera F, Martin-Cuadrado AB, Rodriguez-Brito B, Pasic L, Thingstad TF, Rohwer F, Mira A. Explaining microbial population genomics through phage predation. Nat Rev Microbiol. 2009;7(11):828-36.

37. Kieft K, Zhou Z, Anantharaman K. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. Microbiome. 2020;8(1):90.

38. Kolfenbach JR, Deane KD, Derber LA, O'Donnell C, Weisman MH, Buckner JH, Gersuk VH, Wei S, Mikuls TR, O'Dell J, Gregersen PK, Keating RM, Norris JM, Holers VM. A prospective approach to investigating the natural history of preclinical rheumatoid arthritis (RA) using first-degree relatives of probands with RA. Arthritis Rheum. 2009;61(12):1735-42.

39. Demoruelle MK, Parish MC, Derber LA, Kolfenbach JR, Hughes-Austin JM, Weisman MH, Gilliland W, Edison JD, Buckner JH, Mikuls TR, O'Dell JR, Keating RM, Gregersen PK, Norris JM, Holers VM, Deane KD. Performance of anti-cyclic citrullinated peptide assays differs in subjects at increased risk of rheumatoid arthritis and subjects with established disease. Arthritis Rheum. 2013;65(9):2243-52.

40. Kang DW, Adams JB, Gregory AC, Borody T, Chittick L, Fasano A, Khoruts A, Geis E, Maldonado J, McDonough-Means S, Pollard EL, Roux S, Sadowsky MJ, Lipson KS, Sullivan MB, Caporaso JG, Krajmalnik-Brown R. Microbiota Transfer Therapy alters gut ecosystem and improves gastrointestinal and autism symptoms: an open-label study. Microbiome. 2017;5(1):10.

41. Paez-Espino D, Pavlopoulos GA, Ivanova NN, Kyrpides NC. Nontargeted virus sequence discovery pipeline and virus clustering for metagenomic data. Nat Protoc. 2017;12(8):1673-82.

42. Nayfach S, Camargo AP, Eloe-Fadrosh E, Roux S, Kyrpides N. CheckV: assessing the quality of metagenome-assmebled viral genomes. bioRxiv preprint. 2020.

43. Roux S, Adriaenssens EM, Dutilh BE, Koonin EV, Kropinski AM, Krupovic M, Kuhn JH, Lavigne R, Brister JR, Varsani A, Amid C, Aziz RK, Bordenstein SR, Bork P, Breitbart M, Cochrane GR, Daly RA, Desnues C, Duhaime MB, Emerson JB, Enault F, Fuhrman JA, Hingamp P, Hugenholtz P, Hurwitz BL, Ivanova NN, Labonte JM, Lee KB, Malmstrom RR, Martinez-Garcia M, Mizrachi IK, Ogata H, Paez-Espino D, Petit MA, Putonti C, Rattei T, Reyes A, Rodriguez-Valera F, Rosario K, Schriml L, Schulz F, Steward GF, Sullivan MB, Sunagawa S, Suttle CA, Temperton B, Tringe SG, Thurber RV, Webster NS, Whiteson KL, Wilhelm SW, Wommack KE, Woyke T, Wrighton KC, Yilmaz P, Yoshida T, Young MJ, Yutin N, Allen LZ, Kyrpides NC, Eloe-Fadrosh EA. Minimum Information about an Uncultivated Virus Genome (MIUViG). Nat Biotechnol. 2019;37(1):29-37.

44. Roux S, Paez-Espino D, Chen IA, Palaniappan K, Ratner A, Chu K, Reddy TBK, Nayfach S, Schulz F, Call L, Neches RY, Woyke T, Ivanova NN, Eloe-Fadrosh EA, Kyrpides NC. IMG/VR v3: an integrated ecological and evolutionary framework for interrogating genomes of uncultivated viruses. Nucleic Acids Res. 2020.

45. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, Horvath P. CRISPR provides acquired resistance against viruses in prokaryotes. Science. 2007;315(5819):1709-12.

46. Stern A, Mick E, Tirosh I, Sagy O, Sorek R. CRISPR targeting reveals a reservoir of common phages associated with the human gut microbiome. Genome Res. 2012;22(10):1985-94.

47. Paez-Espino D, Eloe-Fadrosh EA, Pavlopoulos GA, Thomas AD, Huntemann M, Mikhailova N, Rubin E, Ivanova NN, Kyrpides NC. Uncovering Earth's virome. Nature. 2016;536(7617):425-30.

48. Moreno-Gallego JL, Chou SP, Di Rienzi SC, Goodrich JK, Spector TD, Bell JT, Youngblut ND, Hewson I, Reyes A, Ley RE. Virome diversity correlates with intestinal microbiome diversity in adult monozygotic twins. Cell Host Microbe. 2019;25(2):261-72 e5.

49. Roux S, Emerson JB, Eloe-Fadrosh EA, Sullivan MB. Benchmarking viromics: an in silico evaluation of metagenome-enabled estimates of viral community composition and diversity. PeerJ. 2017;5:e3817.

50. Liang G, Zhao C, Zhang H, Mattei L, Sherrill-Mix S, Bittinger K, Kessler LR, Wu GD, Baldassano RN, DeRusso P, Ford E, Elovitz MA, Kelly MS, Patel MZ, Mazhani T, Gerber JS, Kelly A, Zemel BS, Bushman FD. The stepwise assembly of the neonatal virome is modulated by breastfeeding. Nature. 2020;581(7809):470-4.

51. Mukherjee S, Stamatis D, Bertsch J, Ovchinnikova G, Katta HY, Mojica A, Chen IA, Kyrpides NC, Reddy T. Genomes OnLine database (GOLD) v.7: updates and new features. Nucleic Acids Res. 2019;47(D1):D649-D59.

52. Nayfach S, Roux S, Seshadri R, Udwary D, Varghese N, Schulz F, Wu D, Paez-Espino D, Chen IM, Huntemann M, Palaniappan K, Ladau J, Mukherjee S, Reddy TBK, Nielsen T, Kirton E, Faria JP, Edirisinghe JN, Henry CS, Jungbluth SP, Chivian D, Dehal P, Wood-Charlson EM, Arkin AP, Tringe SG, Visel A, Consortium IMD, Woyke T, Mouncey NJ, Ivanova NN, Kyrpides NC, Eloe-Fadrosh EA. A genomic catalog of Earth's microbiomes. Nat Biotechnol. 2020.

53. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.

54. Breitbart M, Thompson LR, Suttle CA, Sullivan MB. Exploring the vast diversity of marine viruses. Oceanography. 2007;20(2):135-9.

55. Thompson LR, Zeng Q, Kelly L, Huang KH, Singer AU, Stubbe J, Chisholm SW. Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism. Proc Natl Acad Sci U S A. 2011;108(39):E757-64.

56. Breitbart M, Bonnain C, Malki K, Sawaya NA. Phage puppet masters of the marine microbial realm. Nat Microbiol. 2018;3(7):754-66.

57. Reyes A, Haynes M, Hanson N, Angly FE, Heath AC, Rohwer F, Gordon JI. Viruses in the faecal microbiota of monozygotic twins and their mothers. Nature. 2010;466(7304):334-8.

58. Chevallereau A, Blasdel BG, De Smet J, Monot M, Zimmermann M, Kogadeeva M, Sauer U, Jorth P, Whiteley M, Debarbieux L, Lavigne R. Next-generation "-omics" approaches reveal a massive alteration of host RNA metabolism during bacteriophage infection of *Pseudomonas aeruginosa*. PLoS Genet. 2016;12(7):e1006134.

59. De Smet J, Zimmermann M, Kogadeeva M, Ceyssens PJ, Vermaelen W, Blasdel B, Bin Jang H, Sauer U, Lavigne R. High coverage metabolomics analysis reveals phage-specific alterations to *Pseudomonas aeruginosa* physiology during infection. ISME J. 2016;10(8):1823-35.

60. Chatterjee A, Willett JLE, Nguyen UT, Monogue B, Palmer KL, Dunny GM, Duerkop BA. Parallel Genomics Uncover Novel Enterococcal-Bacteriophage Interactions. mBio. 2020;11(2).

61. Howard-Varona C, Lindback MM, Bastien GE, Solonenko N, Zayed AA, Jang H, Andreopoulos B, Brewer HM, Glavina Del Rio T, Adkins JN, Paul S, Sullivan MB, Duhaime MB. Phage-specific metabolic reprogramming of virocells. ISME J. 2020;14(4):881-95.

62. Valvano MA, Messner P, Kosma P. Novel pathways for biosynthesis of nucleotide-activated glycero-manno-heptose precursors of bacterial glycoproteins and cell surface polysaccharides. Microbiology (Reading). 2002;148(Pt 7):1979-89.

63. Nakao R, Senpuku H, Watanabe H. *Porphyromonas gingivalis galE* is involved in lipopolysaccharide O-antigen synthesis and biofilm formation. Infect Immun. 2006;74(11):6145-53.

64. Crummett LT, Puxty RJ, Weihe C, Marston MF, Martiny JBH. The genomic content and context of auxiliary metabolic genes in marine cyanomyoviruses. Virology. 2016;499:219-29.

65. Lopez-Perez M, Haro-Moreno JM, de la Torre JR, Rodriguez-Valera F. Novel caudovirales associated with marine group I Thaumarchaeota assembled from metagenomes. Environ Microbiol. 2019;21(6):1980-8.

66. Ignacio-Espinoza JC, Sullivan MB. Phylogenomics of T4 cyanophages: lateral gene transfer in the 'core' and origins of host genes. Environ Microbiol. 2012;14(8):2113-26.

67. Breitbart M. Marine viruses: truth or dare. Ann Rev Mar Sci. 2012;4:425-48.

68. Markenson JA. Worldwide trends in the socioeconomic impact and long-term prognosis of rheumatoid arthritis. Semin Arthritis Rheum. 1991;21(2 Suppl 1):4-12.

69. Hunter TM, Boytsov NN, Zhang X, Schroeder K, Michaud K, Araujo AB. Prevalence of rheumatoid arthritis in the United States adult population in healthcare claims databases, 2004-2014. Rheumatol Int. 2017;37(9):1551-7.

70. Scher JU, Abramson SB. The microbiome and rheumatoid arthritis. Nat Rev Rheumatol. 2011;7(10):569-78.

71. Mirzaei MK, Maurice CF. Menage a trois in the human gut: interactions between host, bacteria and phages. Nat Rev Microbiol. 2017;15(7):397-408.

72. Bin Jang H, Bolduc B, Zablocki O, Kuhn JH, Roux S, Adriaenssens EM, Brister JR, Kropinski AM, Krupovic M, Lavigne R, Turner D, Sullivan MB. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. Nat Biotechnol. 2019;37(6):632-9.

73. Sutton TDS, Hill C. Gut bacteriophage: current understanding and challenges. Front Endocrinol (Lausanne). 2019;10:784.

74. Roux S, Hallam SJ, Woyke T, Sullivan MB. Viral dark matter and virus-host interactions resolved from publicly available microbial genomes. Elife. 2015;4.

75. Liu X, Zeng B, Zhang J, Li W, Mou F, Wang H, Zou Q, Zhong B, Wu L, Wei H, Fang Y. Role of the gut microbiome in modulating arthritis progression in mice. Sci Rep. 2016;6:30594.

76. Jubair WK, Hendrickson JD, Severs EL, Schulz HM, Adhikari S, Ir D, Pagan JD, Anthony RM, Robertson CE, Frank DN, Banda NK, Kuhn KA. Modulation of inflammatory arthritis in mice by gut microbiota through mucosal inflammation andautoantibody generation. Arthritis Rheumatol. 2018;70(8):1220-33.

77. Braschi E, Shojania K, Allan GM. Anti-CCP: a truly helpful rheumatoid arthritis test? Can Fam Physician. 2016;62(3):234.

78. Krych L, Nielsen DS, Hansen AK, Hansen CH. Gut microbial markers are associated with diabetes onset, regulatory imbalance, and IFN-gamma level in NOD mice. Gut Microbes. 2015;6(2):101-9.

79. Chriswell ME, Kuhn KA. Microbiota-mediated mucosal inflammation in arthritis. Best Pract Res Clin Rheumatol. 2019;33(6):101492.

80. Studer N, Desharnais L, Beutler M, Brugiroux S, Terrazos MA, Menin L, Schurch CM, McCoy KD, Kuehne SA, Minton NP, Stecher B, Bernier-Latmani R, Hapfelmeier S. Functional intestinal bile acid 7alpha-dehydroxylation by *Clostridium scindens* associated with protection from *Clostridium difficile* infection in a gnotobiotic mouse model. Front Cell Infect Microbiol. 2016;6:191.

81. Buffie CG, Bucci V, Stein RR, McKenney PT, Ling L, Gobourne A, No D, Liu H, Kinnebrew M, Viale A, Littmann E, van den Brink MR, Jenq RR, Taur Y, Sander C, Cross JR, Toussaint NC, Xavier JB, Pamer EG. Precision microbiome reconstitution restores bile acid mediated resistance to *Clostridium difficile*. Nature. 2015;517(7533):205-8.

82. Duboc H, Rajca S, Rainteau D, Benarous D, Maubert MA, Quervain E, Thomas G, Barbu V, Humbert L, Despras G, Bridonneau C, Dumetz F, Grill JP, Masliah J, Beaugerie L, Cosnes J, Chazouilleres O, Poupon R, Wolf C, Mallet JM, Langella P, Trugnan G, Sokol H, Seksik P. Connecting dysbiosis, bile-acid dysmetabolism and gut inflammation in inflammatory bowel diseases. Gut. 2013;62(4):531-9.

83. Atarashi K, Tanoue T, Oshima K, Suda W, Nagano Y, Nishikawa H, Fukuda S, Saito T, Narushima S, Hase K, Kim S, Fritz JV, Wilmes P, Ueha S, Matsushima K, Ohno H, Olle B, Sakaguchi S, Taniguchi T, Morita H, Hattori M, Honda K. Treg induction by a rationally selected mixture of Clostridia strains from the human microbiota. Nature. 2013;500(7461):232-6.

84. Gerard P. Metabolism of cholesterol and bile acids by the gut microbiota. Pathogens. 2013;3(1):14-24.

85. Yao L, Seaton SC, Ndousse-Fetter S, Adhikari AA, DiBenedetto N, Mina AI, Banks AS, Bry L, Devlin AS. A selective gut bacterial bile salt hydrolase alters host metabolism. Elife. 2018;7.

86. Campbell DE, Ly LK, Ridlon JM, Hsiao A, Whitaker RJ, Degnan PH. Infection with *Bacteroides* phage BV01 alters the host transcriptome and bile acid metabolism in a common human gut microbe. Cell Rep. 2020;32(11):108142.

87. Lozupone CA, Stombaugh JI, Gordon JI, Jansson JK, Knight R. Diversity, stability and resilience of the human gut microbiota. Nature. 2012;489(7415):220-30.

88. Zhu W, Winter MG, Byndloss MX, Spiga L, Duerkop BA, Hughes ER, Buttner L, de Lima Romao E, Behrendt CL, Lopez CA, Sifuentes-Dominguez L, Huff-Hardy K, Wilson RP, Gillis CC, Tukel C, Koh AY, Burstein E, Hooper LV, Baumler AJ, Winter SE. Precision editing of the gut microbiota ameliorates colitis. Nature. 2018;553(7687):208-11.

89. He SM, Wathier M, Podzelinska K, Wong M, McSorley FR, Asfaw A, Hove-Jensen B, Jia Z, Zechel DL. Structure and mechanism of PhnP, a phosphodiesterase of the carbon-phosphorus lyase pathway. Biochemistry. 2011;50(40):8603-15.

90. Martinez A, Ventouras LA, Wilson ST, Karl DM, Delong EF. Metatranscriptomic and functional metagenomic analysis of methylphosphonate utilization by marine bacteria. Front Microbiol. 2013;4:340.

91. Lee KS, Metcalf WW, Wanner BL. Evidence for two phosphonate degradative pathways in *Enterobacter aerogenes*. J Bacteriol. 1992;174(8):2501-10.

92. Artacho A, Isaac S, Nayak R, Flor-Duro A, Alexander M, Koo I, Manasson J, Smith PB, Rosenthal P, Homsi Y, Gulko P, Pons J, Puchades-Carrasco L, Izmirly P, Patterson A, Abramson SB, Pineda-Lucena A, Turnbaugh PJ, Ubeda C, Scher JU. The pre-treatment gut microbiome is associated with lack of response to methotrexate in new onset rheumatoid arthritis. Arthritis Rheumatol. 2020.

93. Hsu BB, Gibson TE, Yeliseyev V, Liu Q, Lyon L, Bry L, Silver PA, Gerber GK. Dynamic modulation of the gut microbiota and metabolome by bacteriophages in a mouse model. Cell Host Microbe. 2019;25(6):803-14 e5.

94. Kuller LH, Mackey RH, Walitt BT, Deane KD, Holers VM, Robinson WH, Sokolove J, Chang Y, Liu S, Parks CG, Wright NC, Moreland LW. Determinants of mortality among postmenopausal women in the women's health initiative who report rheumatoid arthritis. Arthritis Rheumatol. 2014;66(3):497-507.

95. Choi S, Lee KH. Clinical management of seronegative and seropositive rheumatoid arthritis: A comparative study. PLoS One. 2018;13(4):e0195550.

96. Chatterjee A, Duerkop BA. Beyond bacteria: bacteriophage-eukaryotic host interactions reveal emerging paradigms of health and disease. Front Microbiol. 2018;9:1394.

97. Ayyappan P, Harms RZ, Seifert JA, Bemis EA, Feser ML, Deane KD, Demoruelle MK, Mikuls TR, Holers VM, Sarvetnick NE. Heightened levels of antimicrobial response factors in patients with rheumatoid arthritis. Front Immunol. 2020;11:427.

98. Kleiner M, Hooper LV, Duerkop BA. Evaluation of methods to purify virus-like particles for metagenomic sequencing of intestinal viromes. BMC Genomics. 2015;16:7.

99. Duerkop BA, Huo W, Bhardwaj P, Palmer KL, Hooper LV. Molecular basis for lytic bacteriophage resistance in Enterococci. mBio. 2016;7(4).

100. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, Fierer N, Knight R. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proc Natl Acad Sci U S A. 2011;108 Suppl 1:4516-22.

101. Dowd SE, Sun Y, Wolcott RD, Domingo A, Carroll JA. Bacterial tag-encoded FLX amplicon pyrosequencing (bTEFAP) for microbiome studies: bacterial diversity in the ileum of newly weaned *Salmonella*-infected pigs. Foodborne Pathog Dis. 2008;5(4):459-72.

102. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, Weber CF. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol. 2009;75(23):7537-41.

103. Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. Appl Environ Microbiol. 2013;79(17):5112-20.

104. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol. 2006;72(7):5069-72.

105. Bushnell B. BBMap short read aligner, and other bioinformatic tools. 38.56 ed. https://sourceforge.net/projects/bbmap/2019.

106. Peng Y, Leung HC, Yiu SM, Chin FY. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics. 2012;28(11):1420-8.

107. Li D, Luo R, Liu CM, Leung CM, Ting HF, Sadakane K, Yamashita H, Lam TW. MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. Methods. 2016;102:3-11.

Figure 1

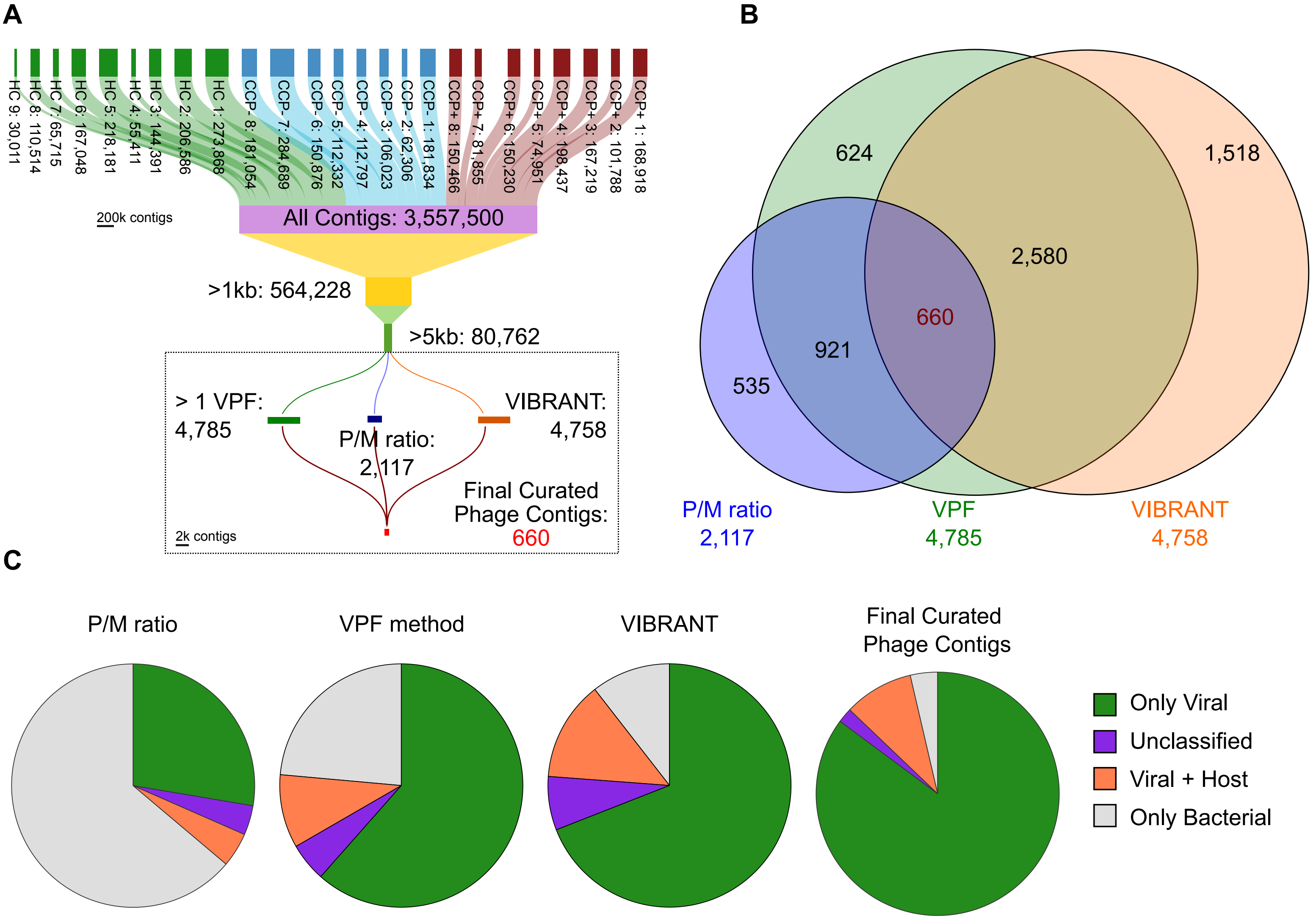

A
HC 9: 30,011
HC 8: 110,514
HC 7: 65,715
HC 6: 167,048
HC 5: 218,181
HC 4: 55,411
HC 3: 144,391
HC 2: 206,586
HC 1: 273,868
CCP- 8: 181,054
CCP- 7: 284,689
CCP- 6: 150,876
CCP- 5: 112,332
CCP- 4: 112,797
CCP- 3: 106,023
CCP- 2: 62,306
CCP- 1: 181,834
CCP+ 8: 150,466
CCP+ 7: 81,855
CCP+ 6: 150,230
CCP+ 5: 74,951
CCP+ 4: 198,437
CCP+ 3: 167,219
CCP+ 2: 101,788
CCP+ 1: 168,918
200k contigs
All Contigs: 3,557,500
>1kb: 564,228
>5kb: 80,762
> 1 VPF:
4,785
P/M ratio:
2,117
VIBRANT:
4,758
Final Curated
Phage Contigs:
660
2k contigs
B
624
1,518
2,580
660
921
535
P/M ratio
2,117
VPF
4,785
VIBRANT
4,758
C
P/M ratio
VPF method
VIBRANT
Final Curated
Phage Contigs
Only Viral
Unclassified
Viral + Host
Only Bacterial

Figure 2

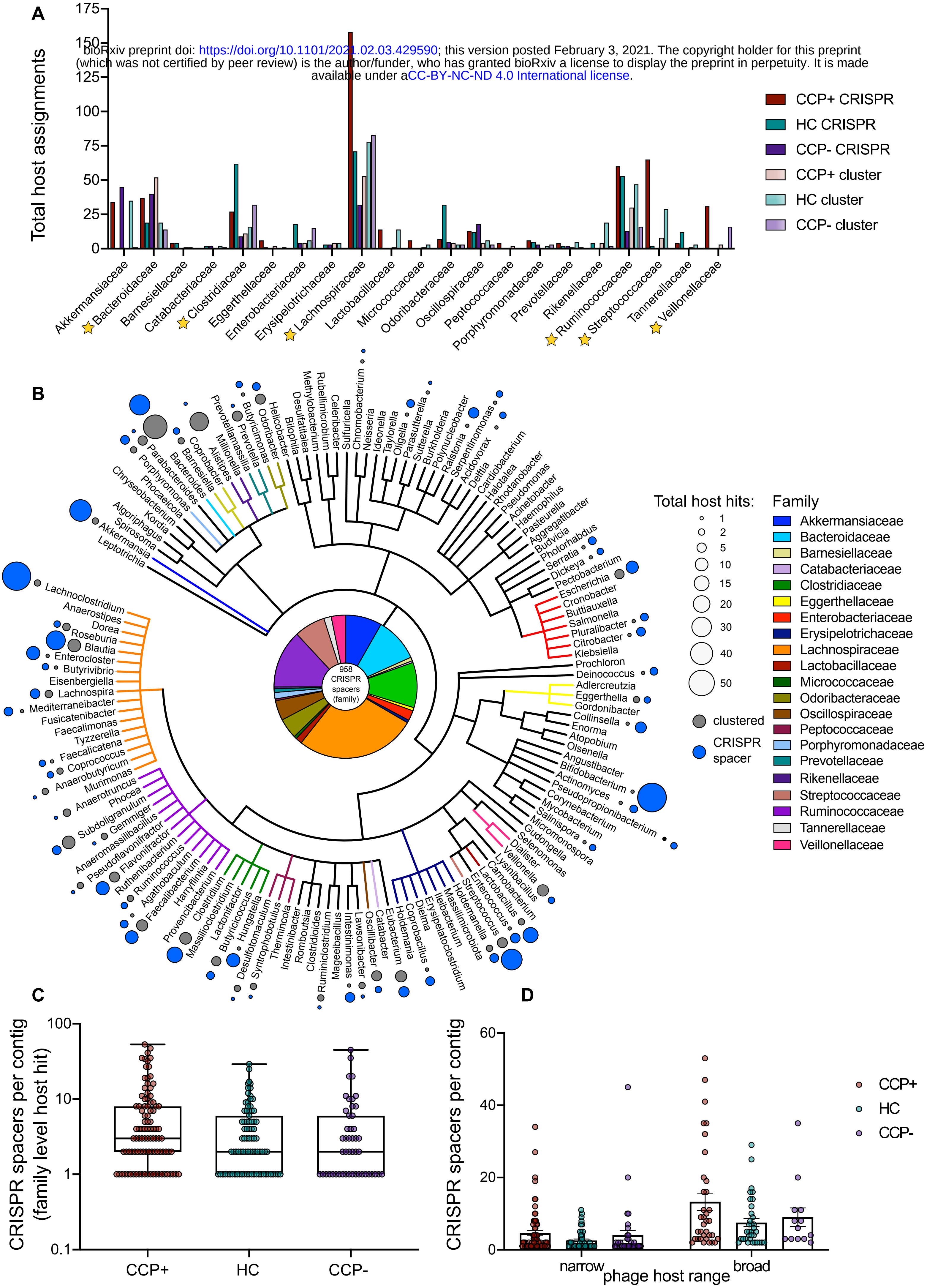

A
Total host assignments
0
25
50
75
100
125
150
175
CCP+ CRISPR
HC CRISPR
CCP- CRISPR
CCP+ cluster
HC cluster
CCP- cluster
Akkermansiaceae
Bacteroidaceae
Barnesiellaceae
Catabacteriaceae
Clostridiaceae
Eggerthellaceae
Enterobacteriaceae
Erysipelotrichaceae
Lachnospiraceae
Lactobacillaceae
Micrococcaceae
Odoribacteraceae
Oscillospiraceae
Peptococcaceae
Porphyromonadaceae
Prevotellaceae
Rikenellaceae
Ruminococcaceae
Streptococcaceae
Tannerellaceae
Veillonellaceae
B
958 CRISPR spacers (family)
Total host hits:
1
2
5
10
15
20
30
40
50
clustered
CRISPR spacer
Family
Akkermansiaceae
Bacteroidaceae
Barnesiellaceae
Catabacteriaceae
Clostridiaceae
Eggerthellaceae
Enterobacteriaceae
Erysipelotrichaceae
Lachnospiraceae
Lactobacillaceae
Micrococcaceae
Odoribacteraceae
Oscillospiraceae
Peptococcaceae
Porphyromonadaceae
Prevotellaceae
Rikenellaceae
Streptococcaceae
Ruminococcaceae
Tannerellaceae
Veillonellaceae
Leptotrichia
Akkermansia
Spirosoma
Algoriphagus
Kordia
Chryseobacterium
Phocaeicola
Porphyromonas
Parabacteroides
Bacteroides
Barnesiella
Coprobacter
Alistipes
Millionella
Prevotellamassilia
Prevotella
Butyricimonas
Odoribacter
Helicobacter
Bilophila
Desulfatitalea
Methylobacterium
Rubellimicrobium
Celeribacter
Sulfuricella
Chromobacterium
Neisseria
Ideonella
Taylorella
Oligella
Parasutterella
Sutterella
Burkholderia
Polynucleobacter
Ralstonia
Serpentinomonas
Acidovorax
Delftia
Cardiobacterium
Halotalea
Rhodanobacter
Pseudomonas
Acinetobacter
Haemophilus
Pasteurella
Aggregatibacter
Budvicia
Photorhabdus
Serratia
Dickeya
Pectobacterium
Escherichia
Cronobacter
Buttiauxella
Salmonella
Pluralibacter
Citrobacter
Klebsiella
Prochloron
Deinococcus
Adlercreutzia
Eggerthella
Gordonibacter
Collinsella
Enorma
Atopobium
Olsenella
Angustibacter
Bifidobacterium
Actinomyces
Pseudopropionibacterium
Corynebacterium
Mycobacterium
Salinispora
Micromonospora
Gudongella
Selenomonas
Dialister
Veillonella
Lysinibacillus
Carnobacterium
Lactobacillus
Enterococcus
Streptococcus
Massilimicrobiota
Holdemanella
Ileibacterium
Erysipelatoclostridium
Dielma
Coprobacillus
Holdemania
Eubacterium
Catabacter
Oscillibacter
Lawsonibacter
Intestinimonas
Mageeibacillus
Ruminiclostridium
Clostridioides
Romboutsia
Intestinibacter
Terrisporobacter
Thermincola
Syntrophobotulus
Desulfotomaculum
Hungatella
Butyricicoccus
Lactonifactor
Massiliociostridium
Clostridium
Provencibacterium
Harryflintia
Faecalibacterium
Agathobaculum
Ruminococcus
Ruthenibacterium
Flavonifractor
Pseudoflavonifractor
Anaeromassilibacillus
Gemmiger
Subdoligranulum
Phocea
Anaerotruncus
Murimonas
Anaerobutyricum
Coprococcus
Faecalicatena
Tyzzerella
Faecalimonas
Fusicatenibacter
Mediterraneibacter
Lachnospira
Eisenbergiella
Butyrivibrio
Enterocloster
Blautia
Roseburia
Dorea
Anaerostipes
Lachnoclostridium
C
CRISPR spacers per contig (family level host hit)
0.1
1
10
100
CCP+
HC
CCP-
D
CRISPR spacers per contig
0
20
40
60
narrow
broad
phage host range
CCP+
HC
CCP-

Figure 3

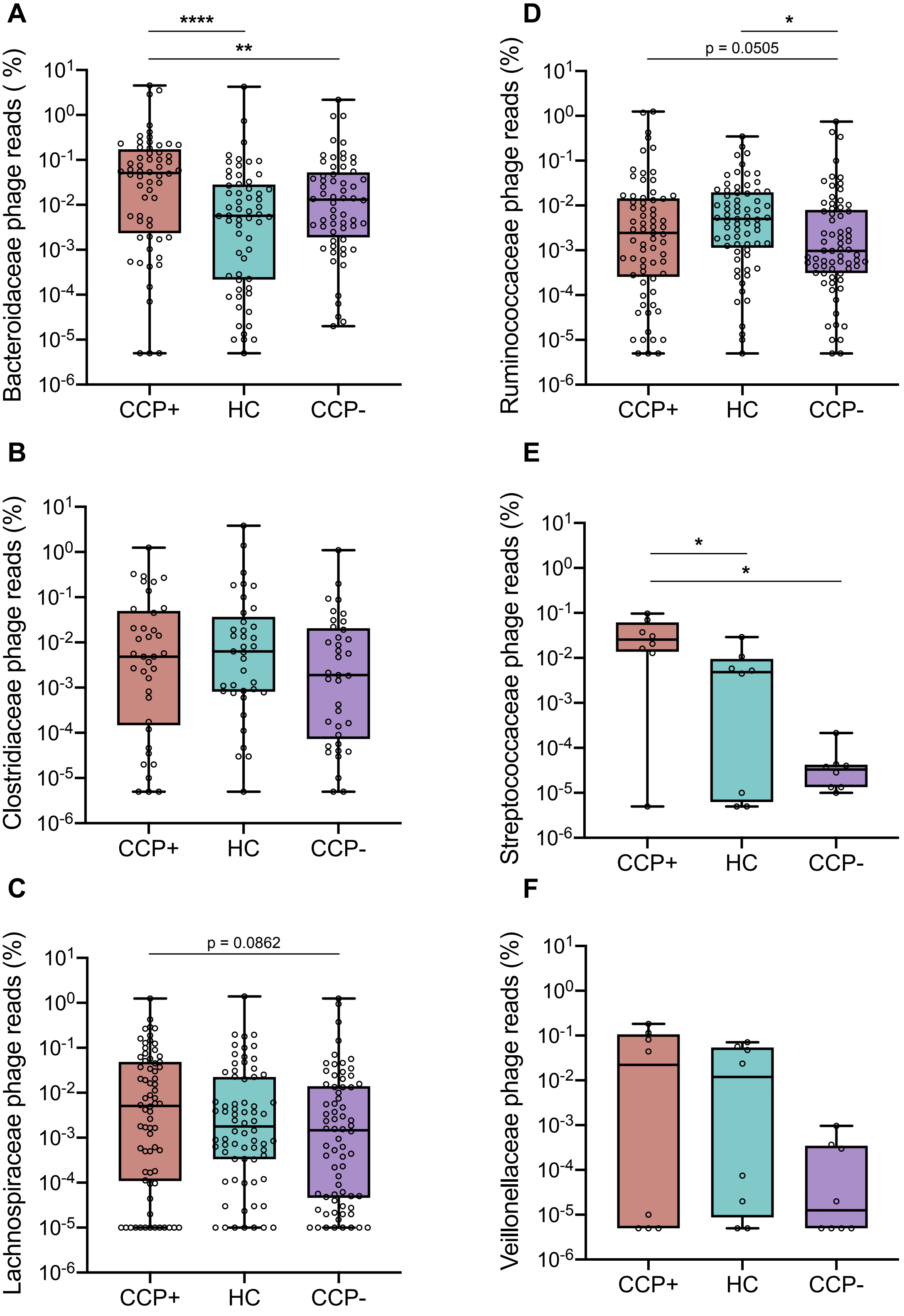

A
****
**
Bacteroidaceae phage reads (%)
10^1
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
CCP+
HC
CCP-
B
Clostridiaceae phage reads (%)
CCP+
HC
CCP-
C
p = 0.0862
Lachnospiraceae phage reads (%)
CCP+
HC
CCP-
D
*
p = 0.0505
Ruminococcaceae phage reads (%)
CCP+
HC
CCP-
E
*
*
Streptococcaceae phage reads (%)
CCP+
HC
CCP-
F
Veillonellaceae phage reads (%)
CCP+
HC
CCP-

Figure 4

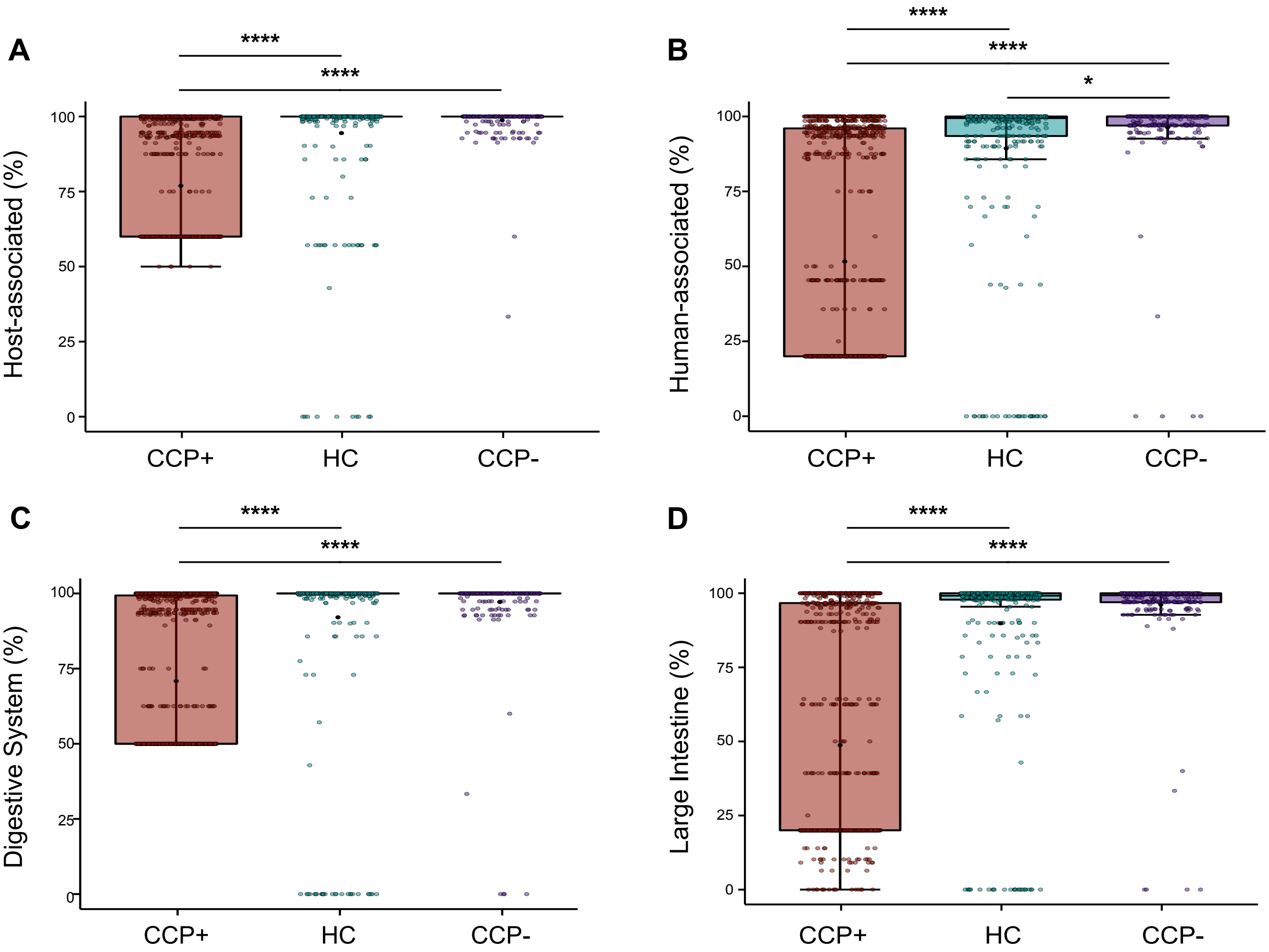

A
****
****
Host-associated (%)
100
75
50
25
0
CCP+
HC
CCP-
B
****
****
*
Human-associated (%)
100
75
50
25
0
CCP+
HC
CCP-
C
****
****
Digestive System (%)
100
75
50
25
0
CCP+
HC
CCP-
D
****
****
Large Intestine (%)
100
75
50
25
0
CCP+
HC
CCP-

Figure 5

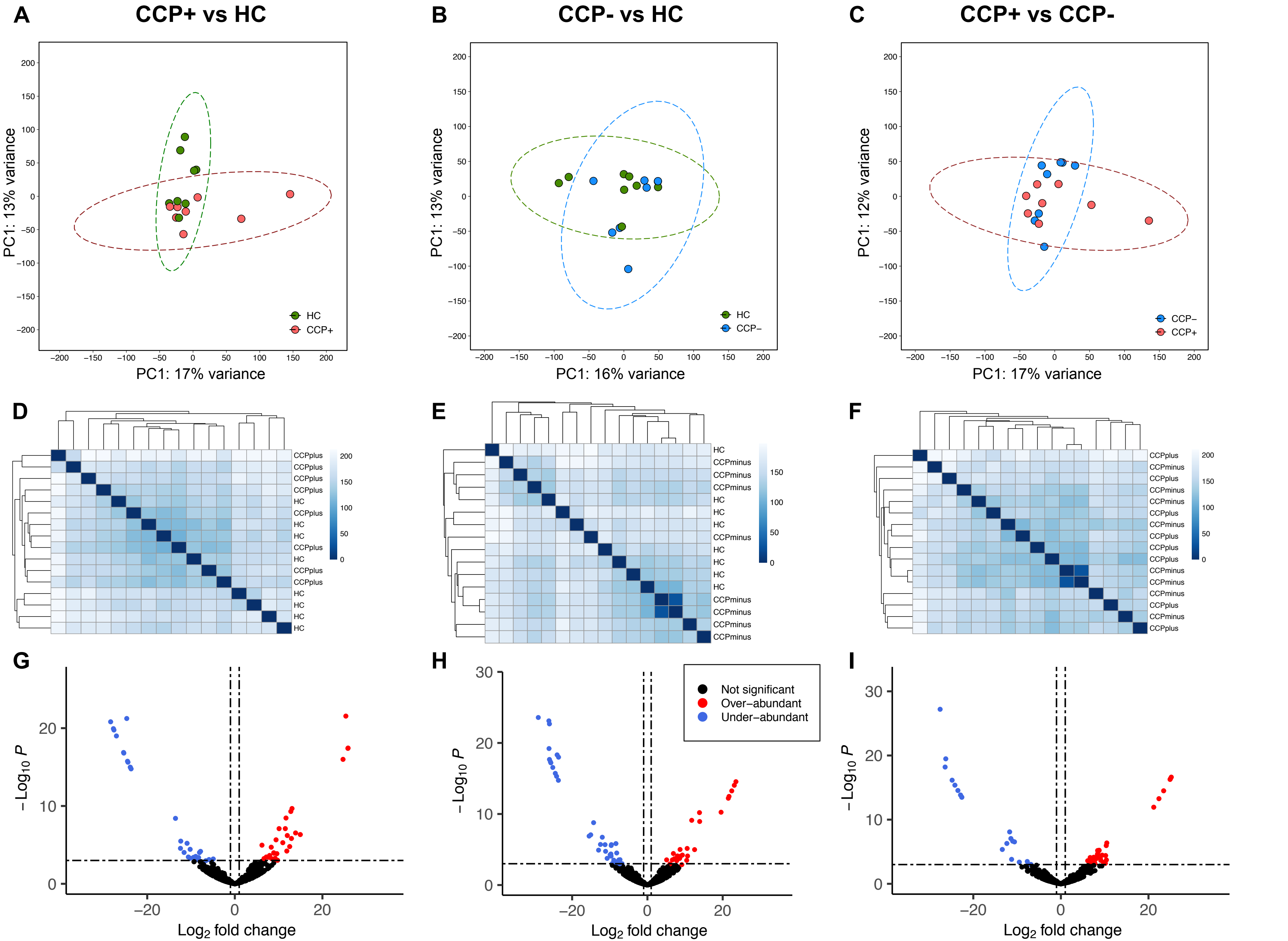
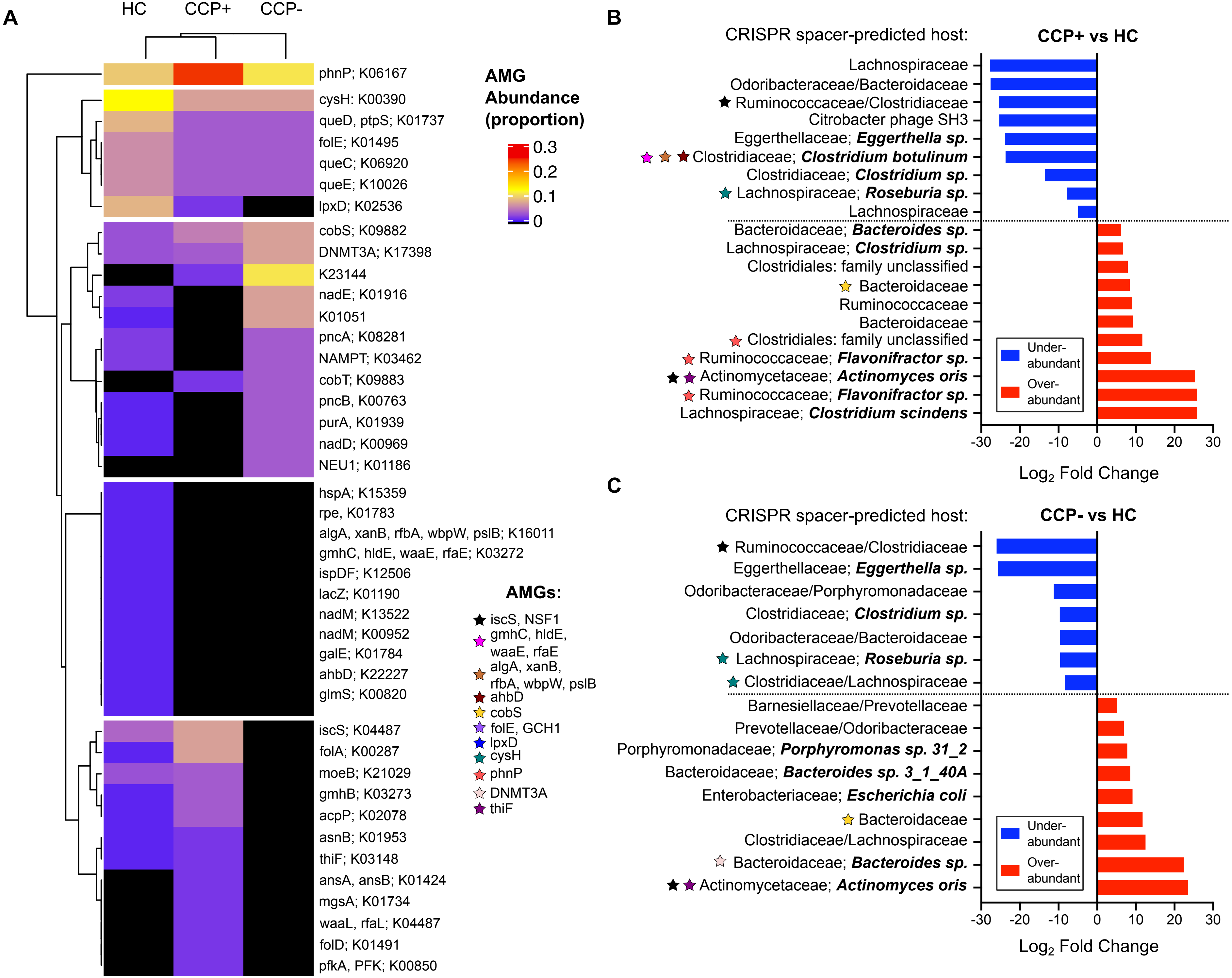

A
CCP+ vs HC
PC1: 13% variance
PC1: 17% variance
HC
CCP+
B
CCP- vs HC
PC1: 13% variance
PC1: 16% variance
HC
CCP−
C
CCP+ vs CCP-
PC1: 12% variance
PC1: 17% variance
CCP−
CCP+
D
CCPplus
CCPplus
CCPplus
CCPplus
HC
CCPplus
HC
HC
CCPplus
HC
CCPplus
CCPplus
HC
HC
HC
HC
E
HC
CCPminus
CCPminus
CCPminus
HC
HC
HC
CCPminus
HC
HC
HC
HC
CCPminus
CCPminus
CCPminus
CCPminus
F
CCPplus
CCPminus
CCPplus
CCPminus
CCPminus
CCPplus
CCPminus
CCPplus
CCPplus
CCPplus
CCPminus
CCPminus
CCPminus
CCPplus
CCPminus
CCPplus
G
H
I
−Log10 P
Log2 fold change
Not significant
Over−abundant
Under−abundant

Figure 6


A
HC
CCP+
CCP-
phnP; K06167
cysH: K00390
queD, ptpS; K01737
folE; K01495
queC; K06920
queE; K10026
lpxD; K02536
cobS; K09882
DNMT3A; K17398
K23144
nadE; K01916
K01051
pncA; K08281
NAMPT; K03462
cobT; K09883
pncB, K00763
purA, K01939
nadD; K00969
NEU1; K01186
hspA; K15359
rpe, K01783
algA, xanB, rfbA, wbpW, pslB; K16011
gmhC, hldE, waaE, rfaE; K03272
ispDF; K12506
lacZ; K01190
nadM; K13522
nadM; K00952
galE; K01784
ahbD; K22227
glmS; K00820
iscS; K04487
folA; K00287
moeB; K21029
gmhB; K03273
acpP; K02078
asnB; K01953
thiF; K03148
ansA, ansB; K01424
mgsA; K01734
waaL, rfaL; K04487
folD; K01491
pfkA, PFK; K00850
AMG Abundance (proportion)
0.3
0.2
0.1
0
AMGs:
★ iscS, NSF1
★ gmhC, hldE, waaE, rfaE
★ algA, xanB, rfbA, wbpW, pslB
★ ahbD
★ cobS
★ folE, GCH1
★ lpxD
★ cysH
★ phnP
☆ DNMT3A
★ thiF
B
CRISPR spacer-predicted host:
CCP+ vs HC
Lachnospiraceae
Odoribacteraceae/Bacteroidaceae
★ Ruminococcaceae/Clostridiaceae
Citrobacter phage SH3
Eggerthellaceae; ***Eggerthella sp.***
★ ★ ★ Clostridiaceae; ***Clostridium botulinum***
Clostridiaceae; ***Clostridium sp.***
★ Lachnospiraceae; ***Roseburia sp.***
Lachnospiraceae
Bacteroidaceae; ***Bacteroides sp.***
Lachnospiraceae; ***Clostridium sp.***
Clostridiales: family unclassified
★ Bacteroidaceae
Ruminococcaceae
Bacteroidaceae
★ Clostridiales: family unclassified
★ Ruminococcaceae; ***Flavonifractor sp.***
★ ★ Actinomycetaceae; ***Actinomyces oris***
★ Ruminococcaceae; ***Flavonifractor sp.***
Lachnospiraceae; ***Clostridium scindens***
Under-abundant
Over-abundant
-30
-20
-10
0
10
20
30
Log$_2$ Fold Change
C
CRISPR spacer-predicted host:
CCP- vs HC
★ Ruminococcaceae/Clostridiaceae
Eggerthellaceae; ***Eggerthella sp.***
Odoribacteraceae/Porphyromonadaceae
Clostridiaceae; ***Clostridium sp.***
Odoribacteraceae/Bacteroidaceae
★ Lachnospiraceae; ***Roseburia sp.***
★ Clostridiaceae/Lachnospiraceae
Barnesiellaceae/Prevotellaceae
Prevotellaceae/Odoribacteraceae
Porphyromonadaceae; ***Porphyromonas sp. 31_2***
Bacteroidaceae; ***Bacteroides sp. 3_1_40A***
Enterobacteriaceae; ***Escherichia coli***
★ Bacteroidaceae
Clostridiaceae/Lachnospiraceae
☆ Bacteroidaceae; ***Bacteroides sp.***
★ ★ Actinomycetaceae; ***Actinomyces oris***
Under-abundant
Over-abundant
-30
-20
-10
0
10
20
30
Log$_2$ Fold Change