Supplementary Appendix

Methods

Sample size estimation

The main objective of this exploratory study was to investigate the difference in the mucosal virome between healthy subjects and patients with ulcerative colitis (UC). In a previous fecal viome study on healthy individuals and patients with inflammatory bowel disease [1], they found a significant higher abundance of *Caudovirales* in the feces of UC subjects than controls (0.6±0.3 versus 0.3±0.2, mean±SD, in relative abundance). We herein hypothesize that the mucosal *Caudovirales* levels in UC is also higher than in controls, and the relative abundance in UC and control mucosa are 0.6±0.2 and 0.3±0.2 respectively. Based on these assumptions, we estimated that 16 number of subjects in each group would achieve 80 percent power at 5 percent significant level. It follows that our current sample size is of sufficient power to detect this difference.

Mucosal virus-like particles (VLPs) enrichment

Rectal biopsy was digested in 1 mL digest buffer (1 mg/mL Collagenase D, 1U Baseline DNase I, PBS pH7.5) at 37℃ for 1 hour, with intermittent intensive vortex every 20 minutes. Biopsy suspension was then cleared by centrifugation at 5,000 x g for 5 minutes to remove debris and cells. Clarified suspension was passed through one 0.22 μm filter to remove residual host and bacteria cells. Samples were treated with lysozyme (1 mg/ml at 37 ℃ for 30 min) followed by chloroform (0.2x volume at RT for 10 min) to degrade any remaining bacterial and host cell membranes. Non-virus protected DNA was degraded by treatment with a DNase cocktail (10U Tubro DNaseI (Ambion), 1U Baseline zero DNase (Epicenter)) followed by heat inactivation of DNases at 65 ℃ for 10 min. VLPs were lysed (4% SDS plus 38 mg/ml Proteinase K at

56 ℃ for 20 min), treated with CTAB (2.5% CTAB plus 0.5 M NaCl at 65℃ for 10 min), and nucleic acid was extracted with phenol:chloroform pH 8.0 (Invitrogen). The aqueous fraction was washed once with an equal volume of chloroform, purified and concentrated on a column (DNA Clean & Concentrator^{TM}-5, Zymo Research). VLP DNA was amplified for 4 hr using Phi29 polymerase (GenomiPhi V2 kit, GE Healthcare) prior to sequencing. To reduce amplification bias, four independent reactions were performed for each sample and pooled together afterwards.

Sequence Processing and Quality Control

Raw reads were filtered by SOAPnuke (v 1.5.3) (http://soap.genomics.org.cn/) developed by BGI as follows: (i) adaptors removed, (ii) read removed if N base is more than 3% of the read, (iii) read removed if bases with quality low than 20 were more than 40% of read, (iv) all duplicates removed. Human sequences were removed from the quality-trimmed dataset by DeconSeq (*v 0.4.3*) with default parameters and the human reference GRCh38 [2].

De Novo Contig Assembly and Taxonomy Annotation

In view of that only part of the reads can be assembled in metagenomic dataset, all other potential viral reads not able to assemble into contigs are considered to present in very low abundances [3] and not taken into analysis. Hence, we assembled those abundant metagenomic viral contigs and classified viral reads. Contigs were assembled using the IDBA (v 1.1.1) [4], using maximum kmer length 120, with a minimum contig length of 1,000 bp. The assembled contigs were clustered at a 95% identity level using CD-HIT [5] to generate a unique contig consortium. In total, 178,155 unique contigs were assembled.

Open Reading Frame (ORF) were predicted and extracted from contigs using the Glimmer3 toolkit (v 3.02) [6] and a minimum length threshold of 100 amino acids. The translated amino acid sequences of predicted ORFs from the VLP contigs were matched against the standard subset of the standalone entire UniProt TrEMBL database as of June 15, 2018, that contained only virus and phage reference proteins, using blastx ($e<10^{-5}$) as described previously [7]. Each contig was assigned taxonomy based on the most abundant taxa contained within that contig using a voting system as described previously for virus taxonomic assignment at different taxon levels [8, 9]. The voting system first annotated each ORF of a contig of interest with the best-hit virus taxonomy. It then compared all of the taxonomic assignments of the ORFs within the contig of interest, and annotated the contig with the majority ORF assignment. Contigs with less than one ORF per 10kb were not assigned taxonomy as this suggests a contig of only limited similarity [8]. Contigs without a majority ORF taxonomic assignment due to ties of multiple major taxa were assigned as having multiple possible taxonomic annotations. Among the assembled all contigs, 51,241 contigs were considered as viral contig based on ORF binning. To preclude false positive contigs classified as viral contig, all these viral contigs were further queried against NCBI nonredundant nucleotide database using BBSketch with the following settings: ambiguous=radom, qtrim=lr, minid=0.97. Viral contigs mapped to organisms from other kingdoms with *ANI*>0.97, most of which were bacteria, were excluded. Eventually, 30,853 contigs were remained and regarded as *bona fide* viral, which were then used as our curated viral contig database. Some contigs shared the same taxonomic identity, the contig table was therefore collapsed by taxonomic identity, where the contig abundances were summed if they shared identity.

Mucosal virome metacommunity clustering analysis was performed according to the previously reported method [10, 11], based on the virome species-level

community structures. Virus core species, common species and unique species were defined when the viral species meet the below criteria: 1) core species was deemed when it exists in >50% of all study subjects with concordant RPKM>10 in any individual; 2) common species was deemed when it exists in 20-50% of all study subjects with concordant RPKM>10 in any individual; and 3) unique species was deemed when it exists in <20% of all study subjects. The presence and percentage ratio were evaluated in each individual and compared between healthy controls, UC Metacommunity 1 and UC metacommunity 2 subjects.

Non-metric multidimensional scaling analysis

The difference in mucosal viral community structures between controls and UC was performed via NMDS (Non-metric multidimensional scaling analysis) plot based upon Bray-Curtis dissimilarities among all subjects.

DESeq, Random Forest and LEfSe linear discriminant analysis

To compare differences in the configurations of mucosal virome and bacterial microbiome as well as the functions of virome between UC patients and healthy household controls, between mucosal virome metacommunities, Differential analyses were performed. *DESeq* and *Random Forest* were performed in R via DEseq and randomForest package respectively. Lefse analyses were performed on the Huttenhower lab Galaxy server (http://huttenhower.sph.harvard.edu/galaxy/). *DB-RDA* analysis was also performed in R to delineate the effect of medication and health on mucosal virome configutation.

Virome function analysis

Virome functions were classified through annotating all viral-contig derived reads via *HUMANN2 v0.9.4*. Predicted functions were collapsed by Gene ontology terms and Pfam protein family identies, with abundance values expressed in RPK (reads per kilobase). To establish the presence or absence a function within a sample, a stringent RPK threshold value > 10 was used to define as present.

Mucosal bacterial DNA Extraction

Bacterial DNA was extracted from rectal biopsies using Maxwell RSC Tissue DNA kit (Promega, Madison, Wisconsin) according to the protocal. Briefly, 1 rectal biopsy was added to the Lysis Tube with 0.5mm plus 0.1 mm beads filled with 300 µl Lysis solution (500mM NaCl, 50mM Tris-HCl pH 8.0, 50mM EDTA, 4% sodium dodecyl sulphate SDS), and then homogenized at maximum speed for ⩾ 3 minutes. The lysates were incubated at 70℃for 10 minutes followed by centrifugation at ⩾ 16,000 x g for 5 minutes. The supernatant was transferred to Maxwell RSC Tissue tools and extracted through Maxwell RSC Instrument (Promega, Madison, Wisconsin).

16S rRNA sequencing and quality control

The final fecal DNA samples were sequenced on the Illumina Hiseq 2500 platform (V4 region, 2 X 250 bp). Quality control and data analysis were implemented in mothur (v 1.38.0) as previously described [12]. Any sequences with ambiguous bases and anything longer than 275 bp were removed, and aligned against the non-redundant Greengenes database (v 13.8) [13] using the NAST algorithm. Any sequences that failed to align with the V3-4 region were discarded. The remaining sequences were trimmed to the same alignment coordinates over which they fully overlapped, followed by removal of homopolymers and detection for the presence of chimeras by UChime.

16S rRNA sequencing data analysis

The resulting sequences were classified against the Greengenes database and annotated with deepest level taxa represented by pseudo-bootstrap confidence scores of at least 80% averaged over 1,000 iterations of the naive Bayesian classifier. Any sequences that were classified as either being originated from archaea, eukarya, chloroplasts, mitochondria, or unknown kingdoms, were removed. The annotated sequences were assigned to phylotypes according to their consensus taxonomy with which at least 80% of the sequences agreed. Closed reference operational taxonomic units (OTUs) sharing 97% identity were clustered as well and assigned taxonomy according to the Greengenes database. Bacterial diverisy and richness were calculated in R via *phyloseq* package. Bacteria abundance tables at the phylum, family and genus levels were also generated and compared between health and UC mucosa. Lefse analysis was performed to define bacterial taxa associated with UC and healthy controls.

Reference

1 Norman JM, Handley SA, Baldridge MT, Droit L, Liu CY, Keller BC*, et al.* Disease-Specific Alterations in the Enteric Virome in Inflammatory Bowel Disease. Cell 2015;**160**:447-60.

2 Schmieder R, Edwards R. Fast Identification and Removal of Sequence Contamination from Genomic and Metagenomic Datasets. PloS one 2011;**6**.

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

4 Peng Y, Leung HCM, Yiu SM, Chin FYL. IDBA - A Practical Iterative de Bruijn Graph De Novo Assembler. Lect N Bioinformat 2010;**6044**:426-40.

5 Fu LM, Niu BF, Zhu ZW, Wu ST, Li WZ. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 2012;**28**:3150-2.

6 Delcher AL, Bratke KA, Powers EC, Salzberg SL. Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics 2007;**23**:673-9.

7 Hannigan GD, Meisel JS, Tyldsley AS, Zheng Q, Hodkinson BP, SanMiguel AJ*, et al.* The Human Skin Double-Stranded DNA Virome: Topographical and Temporal Diversity, Genetic Enrichment, and Dynamic Associations with the Host Microbiome. mBio 2015;**6**.
8 Minot S, Bryson A, Chehoud C, Wu GD, Lewis JD, Bushman FD. Rapid evolution of the human gut virome. Proceedings of the National Academy of Sciences of the United States of America 2013;**110**:12450-5.
9 Hannigan GD, Meisel JS, Tyldsley AS, Zheng Q, Hodkinson BP, SanMiguel AJ*, et al.* The human skin double-stranded DNA virome: topographical and temporal diversity, genetic enrichment, and dynamic associations with the host microbiome. mBio 2015;**6**:e01578-15.
10 Costea PI, Hildebrand F, Manimozhiyan A, Bäckhed F, Blaser MJ, Bushman FD*, et al.* Enterotypes in the landscape of gut microbial community composition. Nat Microbiol 2018;**3**:8.
11 Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR*, et al.* Enterotypes of the human gut microbiome. Nature 2011;**473**:174.
12 Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB*, et al.* Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. Appl Environ Microb 2009;**75**:7537-41.
13 McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A*, et al.* An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. Isme J 2012;**6**:610-8.

Supplementary Figure 1. Mucosal virobiota composition at the order, genus and species levels in healthy individuals and UC subjects. **a.** Mucosal viral abundance variation boxplot in healthy individuals. **b.** Mucosal viral abundance variation boxplot in UC subjects. Only the most abundant viral taxa are shown for clarity. For all box plots, the boxes extend from the 1st to 3rd quartile (25^{th} to 75^{th} percentile), with the median depicted by a horizontal line.

Supplementary Figure 2. Mucosal virobiota diversity alteration at the species level in UC. Comparison of the mucosal virobiota α diverisity based on Shannon diversity (a) and Chao1 richness (b) index in the mucosa of controls and UC subjects. Statistical significance was determined by *t* test, **P* < 0.05, ***P* < 0.01. For box plots, the boxes extend from the 1st to 3rd quartile (25^{th} to 75^{th} percentile), with the median depicted by a horizontal line.

Supplementary Figure 3. Mucosal inflammation was linked to virome alterations in UC. **a**, Comparison of the virus abundance in the rectal mucosa of controls and UC subjects with and without inflammation, at the viral contig level. Total virus abundance was calculated as RPKM sum of all viral contigs recruited reads normalized by sequence depth in each subject. Between-group difference was determined by Mann-Whitney test, **P* < 0.05. **b**, Comparison of the mucosal virome α diverisity based on Shannon diversity, evenness, Chao1 richness index in the mucosa of controls and UC subjects with and without inflammation. Statistical significance was determined by one-way *ANOVA*, **P* < 0.05, ***P* < 0.01. **c**, Comparison of the *Caudovirales* order abundance in the rectal mucosa of controls and UC subjects with and without inflammation. *Caudovirales* abundance was calculated as RPKM sum of *Caudovirales* contigs recruited reads normalized by sequence depth in each subject. Between-group difference was determined by Mann-Whitney test, **P* < 0.05. **d**, Comparison of the mucosal *Caudovirales* α diverisity based on Shannon diversity, evenness, Chao1 richness index in the mucosa of controls and UC subjects with and without inflammation. Statistical significance was determined by one-way *ANOVA*, **P* < 0.05. For box plots, the boxes extend from the 1st to 3rd quartile (25th to 75th percentile), with the median depicted by a vertical line. Dots indicate individual values of the studied subjects.

Supplementary Figure 4. Phylogentic dendrogram of mucosa viral taxa in health and UC. Differentially enriched viral taxa between healthy and UC mucosa are color-coated, at the order, family and genus levels. Red colored taxa correspond to UC mucosa enriched viruses, while blue colored taxa correspond to healthy individual enriched viruses.

Supplementary Figure 5. Validation of differential viral species in Beijing and Xiangshan cohorts. **a**, Geographic plot of study sites in China. 20 healthy controls versus 20 UC subjects and 8 healthy controls versus 8 UC subjects were enrolled from study sites Bejing and Xiangshan, Zhejiang province, respectively. Abundances of the species *Chrysochromulina ericina virus* (**b**) and *Mimivirus* (**c**) at study sites Beijing and Xiangshan were plotted in mean±s.e.m. Between-group difference was determined by Mann-Whitney test, **P* < 0.05, ***P* < 0.01, **P* < 0.001. Dots indicate individual values of the studied subjects.

Supplementary Figure 6. Differential viral species between mucosal virome Metacommunity 1 subjects in health and UC. **a**, Differentially enriched viral species between the mucosa of healthy Metacommunity 1 subjects and UC Metacommunity 1 subjects, determined by *DESeq* analysis with FDR correction. Only those taxa with adjusted *p* values < 0.05 and $|Log_2$(between-group fold-change)| >2 are shown. Blue denotes viral species enriched in healthy metacommunity 1 subjects, red denotes viral species enriched in UC metacommunity 1 subjects. The abundances of two replicated viral species, *Feldmannia species virus* (**b**) and *Psedomonas virus* (**c**), in Beijing and Xiangshan cohort, were plotted with respect to each study site. Between-group statistical difference was determined by Mann-Whitney test, **P* < 0.05. Dots indicate individual values of the studied subjects.

Supplementary Figure 7. Mucosal virome functions in healthy individuals and UC subjects, predicted at the Gene ontology and Pfam protein family levels. **a.** The abundance boxplot of mucosal viral function in healthy individuals. **b.** The abundance boxplot of mucosal viral function in UC subjects. Only the most abundant 20 viral functions, GO and Pfam protein functions respectively, are shown for clarity. For box plots, the boxes extend from the 1st to 3rd quartile (25^{th} to 75^{th} percentile), with the median depicted by a horizontal line.

Supplementary Figure 8. Altered bacterial microbiota in UC mucosa. **a**, Comparison of bacteria α diversities based on Simpson diversity and Chao1 richness in the mucosa of controls and UC subjects. Statistical significance was determined by *t* test, **P* < 0.05, ***P* < 0.01. For box plots, the boxes extend from the 1st to 3rd quartile (25^{th} to 75^{th} percentile), with the median depicted by a horizontal line. The bacteria composition in health and UC mucosa was plotted in relative abundance, at the phylum (**b**), family (**c**) and genus (**d**) levels. **e**, Differentially enriched bacterial taxa between the mucosa of health individuals and UC subjects were determined by *Lefse* analysis with

FDR correction. Only those taxa with adjusted *P* value < 0.05 and LDA effect size >2 are shown.