Exploring the gut virome in fecal immunochemical test stool samples reveals novel associations with lifestyle in a large population-based study

Paula Istvan[1*], Einar Birkeland[1*], Ekaterina Avershina[2,3], Ane S Kværner[4], Vahid Bemanian[5], Willem M. de Vos[6,7], Torbjørn Rognes[1,8], Paula Berstad[4], Trine B Rounge[2,3,9#]

Affiliations:

1. Centre for Bioinformatics, Department of Informatics, University of Oslo, Norway
2. Department of Tumor Biology, Institute of Cancer Research, Oslo University Hospital, Norway
3. Centre for Bioinformatics, Department of Pharmacy, University of Oslo, Norway
4. Section for Colorectal Cancer Screening, Cancer Registry of Norway
5. Pathology Department, Akershus University Hospital, Norway
6. Human Microbiome Research Program, Faculty of Medicine, University of Helsinki, Finland
7. Laboratory of Microbiology, Wageningen University, The Netherlands
8. Department of Microbiology, Oslo University Hospital, Oslo, Norway
9. Department of Research, Cancer Registry of Norway, Norway

*Equal contribution

#Corresponding author: Trine B Rounge, trinro@uio.no

ABSTRACT

Stool samples for fecal immunochemical tests (FIT) are collected in large numbers worldwide as part of colorectal cancer screening programs, but to our knowledge, the utility of these samples for virome studies is still unexplored. Employing FIT samples from 1034 CRCbiome participants, recruited from a Norwegian colorectal cancer screening study, we identified and annotated more than 18000 virus clusters (vOTUs), using shotgun metagenome sequencing. Only six percent of vOTUs were assigned to a known taxonomic family, with *Microviridae* being the most prevalent viral family. Genome integration state was family-associated, and the majority of identified viruses were unintegrated. Linking individual profiles to comprehensive lifestyle and demographic data showed 17/25 of the variables to be associated with the gut virome. Physical activity, smoking, and dietary fiber consumption exhibited strong and consistent associations with both diversity and relative abundance of individual vOTUs, as well as with enrichment for auxiliary metabolic genes.

We demonstrate the suitability of FIT samples for virome analysis, opening an opportunity for large-scale studies of this yet enigmatic part of the gut microbiome. The diverse viral populations and their connections to the individual lifestyle uncovered herein paves the way for further exploration of the role of the gut virome in health and disease.

INTRODUCTION

Gut residing viruses represent an important component of the intestinal microbial ecosystem and may be collectively referred to as the gut virome. Recent large-scale efforts have shown the virome to comprise a vast and diverse population[1–5], of which bacteriophages (phages), i.e. viruses that infect and replicate in bacteria and archaea, make up the overwhelming majority. However, the extent of virome diversity in the gut remains poorly annotated, with only a minor fraction typically assigned taxonomy[2].

Viruses residing in the human gut are thought to act as a key modulator of the gut microbiome through their interaction with bacteria and the host immune system[6]. They may influence the structure and function of the bacterial community through facilitation of horizontal gene transfer[7], nutrient recycling, regulation of bacterial virulence[8], and gain of antibacterial resistance[9]. Furthermore, viruses play a direct and indirect role in interactions between the human host and the bacterial community[10], and have been shown to exhibit temporal stability as high as that of their bacterial hosts[11,12].

The gut virome has been linked to human host and environmental factors, for specific food items[3,13] or viral populations[14], and like the bacterial community, its composition has been found to develop as a function of age[2]. The gut virome has also been associated with major chronic diseases such as inflammatory bowel disease and type 2 diabetes[15,16]. Dysregulation of gut bacteria and abundance of certain bacteria[17–19] are also proposed features of the association between the gut microbiome and

colorectal cancer development[20]. These changes in the bacteriome are likely to be accompanied by phage dysregulation[21].

Given high diversity and interindividual variability of the gut virome, large population-scale analyses are needed to decipher its role in human health and disease. Colorectal cancer screening programs, inviting millions each year, are currently running or in the planning stages in many countries across the globe[22]. A widely used screening strategy is based on fecal occult blood testing of gut samples, the fecal immunochemical test (FIT). The FIT is non-invasive, inexpensive, and scalable to large populations[23]. There is accumulating evidence that these gut samples are suitable for analysis of various features of the gut microbiome[24–26]. Combining the large numbers of gut samples from population-based screening programs with affordable shotgun metagenomics could propel unbiased and population-based virome studies.

To the best of our knowledge, no studies have yet been conducted analyzing the gut virome using FIT samples. With the availability of a large number of FIT samples collected in a Norwegian colorectal cancer screening trial, we have performed comprehensive profiling of the gut virome. We describe viral diversity including taxonomy, genome integration and functional potential, and assess associations of these factors with individual diet, lifestyle and demographic factors.

MATERIALS AND METHODS

Study population

The CRCbiome enrolled individuals aged 55–76 who tested positive for FIT (and were referred for colonoscopy) from the Bowel Cancer Screening in Norway (BCSN) trial, which is a population-wide randomized trial comparing the effectiveness of once-only sigmoidoscopy and biennial FIT testing. Out of the 2700 individuals invited to participate, 1640 met the inclusion criteria and provided informed consent. Details on recruitment procedures can be found in Kværner et al[27]. All participants provided FIT samples containing fecal matter that were self-collected at home and shipped to the laboratory by mail at ambient temperature. Following FIT testing, samples were stored at -80°C until withdrawal of leftover buffer from the FIT container (~1600 µl; containing about 10 mg fecal matter) and DNA extraction (see details below). For the purpose of the CRCbiome overall aim, samples were selected based on their colonoscopy results, excluding those without colonoscopy, or with findings of uncertain clinical significance. The availability of sufficient DNA (>0.7 ng/µl) and metagenome data (>1 gigabase after QC) was also required. The final number of FIT metagenomes included in the study was 1034 (Fig. 1a).

Questionnaire data

Prior to the colonoscopy, participants were asked to complete two questionnaires on diet, lifestyle, and demography: a food frequency questionnaire (FFQ), developed and validated[28–30] at the Department of Nutrition, University of Oslo, and a lifestyle and demographic questionnaire (LDQ), developed in-house. The FFQ is designed to capture the habitual diet during the preceding year. The current questionnaire version includes a total of 23 questions, covering 256 food items. For each food item, participants were asked to record frequency of consumption, ranging from never/seldom to several times a day, and/or amount, typically as portion sizes given in various household units. Dietary intake was calculated using the food and nutrient calculation system, KBS, developed at the Department of Nutrition, University of Oslo, with its associated database, which is largely based on the Norwegian Food Composition Table[31]. We focused on key dietary measures, including total energy intake (kcal/day), intake of macronutrients (in g/day or energy percentage (E%)), and selected food groups (g/day), being linked to risk of major chronic diseases such as cancer (described in further detail below)[32,33]. The FFQ also included questions on body weight (kg) and height (m), which was used to calculate participants' BMI (kg/m^2). The LDQ is a questionnaire developed specifically for the CRCbiome study to obtain data on key lifestyle and demographic variables. The questionnaire includes ten questions in total, where the ones relevant to the current study included demographic factors (national background, education, occupation and marital status), antibiotic and antacid usage during the last three months, smoking and snus habits, and physical activity level. In the question concerning tobacco usage, participants were asked about their current habits, including the daily number of cigarettes/snus portions, and to recall years since possible cessation and total years of use. In the present study, smokers and snusers were defined as self-reported regular or occasional users, or those being registered with recent use (<10 years). For physical activity, participants were asked to report the time spent in low, moderate and vigorous physical activity per week during the past year. Total amount of moderate to vigorous physical activity (min/week) was calculated by summing the time spent in moderate and vigorous activity, the latter weighted by a factor of two to best match national[34] and international recommendations[35,36].

As a measure of the overall diet and lifestyle, we created a healthy lifestyle index (HLI), grading participants by adherence to the following seven recommendations (primary intended to prevent cancer, but is also relevant for other major chronic disease): 1) be a healthy body weight, 2) be physically active, 3) consume a diet rich in whole grains, vegetables, fruit, and beans, 4) limit intake of "fast foods" and other processed foods high in fat, starches, or sugars, 5) limit consumption of red and processed meat, 6) limit consumption of sugar sweetened drinks, and 7) limit alcohol consumption. Further details on the HLI can be found Kværner et al.[37].

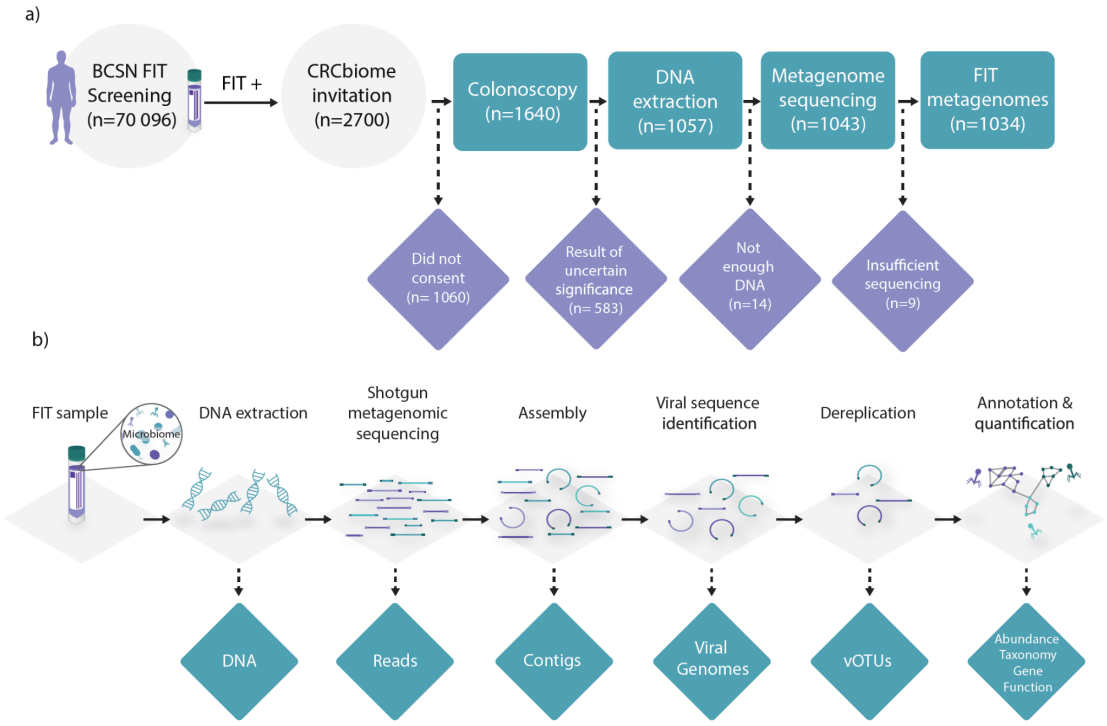

Fig. 1 Study design. a) participant flowchart. 2700 FIT positive Bowel Cancer Screening in Norway (BCSN) participants were invited to the study. Excluded samples are indicated in purple. *Participants were excluded if they had findings of uncertain clinical significance, i.e., a low number of non-advanced adenomas or non-advanced sessile serrated lesions. b) Workflow for virome characterization. DNA was extracted from the FIT leftover buffer. Shotgun metagenomic sequencing was performed on the Illumina platform and the resulting reads were assembled using metaSPAdes. Viral genomes were identified using Virsorter2, and then dereplicated using Galah. Representative vOTUs were taxonomically annotated using vConTACT2. DRAMv was used for annotation of gene function. For details, see materials and methods.

Sample collection, library generation and metagenome sequencing

Following collection of FIT sampling kits and measurement of fecal occult blood concentration, leftover buffer was used as input material for DNA extraction and library preparation for the generation of shotgun metagenome sequencing data. DNA was extracted using the QIAsymphony automated extraction system using an off-board lysis protocol described in Kværner et al[27]. Sequencing libraries were constructed according to the Nextera DNA Flex Library Prep Reference Guide, except scaling down the reaction volumes to one quarter of the reference. Library pools of 240 samples were combined and size selected to a fragment size of 650–900 bp. Sequencing was performed on the Illumina NovaSeq system using S4 flow cells with lane divider, with each pool sequenced on a single lane resulting in paired-end 2x151 bp reads. Shotgun metagenome sequencing was performed aiming to achieve 3 gigabases per sample.

Sequence reads quality control and assembly

The metagenome processing framework Metagenome-ATLAS[38] was used for sequencing quality control and assembly. In brief, ATLAS utilizes BBTools[39] utilities for adapter and quality trimming of reads, and for the removal of human genome and PhiX reads. Quality trimmed reads, both paired and unpaired, were used for *de novo* assembly using metaSPAdes[40]. For information on versions of tools and databases used, see Supplementary Table 4.

Viral sequence identification, dereplication, quantification and assessment of genome integration

Viral genomes were classified using VirSorter2[41] with metagenomic contigs >1500 bp as input. CheckV[42] was used for assessment of genome completeness, quality, level of host sequence content, annotation of host genome integration, and to extract the fractions of contigs determined to contain viral sequences. Viral genomes assigned a quality of medium or higher (corresponding to >50% completeness) by CheckV assessment were considered for further analysis. We clustered viral genomes by average nucleotide identity (ANI) to define viral operational taxonomic units, or clusters (vOTUs) using the dereplication tool Galah[43], defining clusters by an ANI threshold of 97% covering at least 70% of each genome's length. The viral genome with the highest completeness in each cluster was chosen as the representative genome for that vOTU. Quality controlled paired-end reads from all participants were mapped to each vOTU using BBMap[39], with the following options: *pairlen=1000*, *pairedonly=t*, *minid=0.9*, *maxindel=100*, *ambiguous=all*, *maxsites=10*. The vOTU coverage was calculated using the *pileup* function from BBTools, and vOTU abundance was recorded as the median coverage for those with reads mapping to at least 75% of the genome.

Annotation of viral genomes

Taxonomic classification of vOTUs was carried out using vConTACT2[44], based on proteins identified with Prodigal[45]. To establish a reference database for classification, INPHARED[46], a tool for automated retrieval and creating of custom databases based on viral protein sequences and associated metadata, updated monthly, was used (June 13th, 2023 update). vConTACT2 uses a network-based approach to identify viral clusters based on clustering of viral proteins. For processing of vConTACT2 clustering, graphanalyzer[47] was used. Here, taxonomy was assigned if a vOTU had a direct or indirect connection (up to one degree removed) to a reference, where the strength of the connection prioritized the taxonomy assignment. vConTACT2 was run with parameters *--db 'ProkaryoticViralRefSeq94-Merged' --rel-mode 'Diamond' --pcs-mode MCL --vcs-mode ClusterONE*. Cytoscape[48] was used to visualize the vOTU network excluding vOTUs with no significant associations (outliers).

DRAM-v[49] was employed for gene annotation of vOTUs, using the databases Pfam[50], VOGDB[51], KOfam[52], UniRef90[53], dbCAN[54] and RefSeq[55]. Auxiliary metabolic genes (AMGs) were defined using default settings in DRAM-v. AMGs were counted by presence/absence of each category of AMG per vOTU. For versions of these tools and databases, see Supplementary Table 4.

Statistics

The R package vegan[56] was used to calculate alpha-diversity (inverse Simpson index), with between-group differences assessed using ANOVA tests, adjusting for sequencing depth. Beta-diversity (Bray-Curtis dissimilarity matrices) and differences between groups were evaluated using PERMANOVA implemented in the "*vegan:adonis2*" function with 999 permutations. Differential abundance of vOTUs was evaluated using the R package MaAsLin2[57] using a linear model with total sum scaling normalization, and adjustment for age group (50-60, 60-70, and 70-80), sex, and geographic region (Bærum and Moss regions, the two recruitment regions in South-East Norway). To examine associations with diet, lifestyle, and demographic variables measured on a continuous scale, variables were grouped into tertiles. Comparisons were then made of virome variables between the lowest and highest tertiles. Participants with missing data or selecting the answer option "Unknown" (applicable to the items concerning antibiotic and antacid usage), were excluded from statistical analyses evaluating associations with diversity, composition and differential abundance. The magnitudes of observed associations with alpha and beta diversity were quantified using Omega-squared statistics[58], which for beta diversity was calculated employing the "*adonis_OmegaSq*" function from the R package micEco. Custom R scripts were used for statistics and visualization of results (https://github.com/Rounge-lab/CRCbiome_virome_2023).

RESULTS

Dataset description

Raw shotgun metagenomic sequencing data comprised 13.5 billion paired end reads, with 11.5 billion passing QC (median of 10.7 million reads per sample, IQR= 3.5 million; Fig. 2a). Storage time of samples before DNA extraction ranged from 34 to 1301 days, with a median of 198 days (Fig. 2a). Storage time did not impact DNA concentration, sequencing depth, assembly quality or the number of retrieved viral genomes ($|rho| \leq 0.05$, Fig. 2c). Spearman's rank correlation of DNA concentration to the sequencing depth, number of retrieved viral genomes and alpha-diversity ranged between rho=0.15 and rho=0.18, whereas correlation to the assembly quality was negligible (rho = 0.04, Fig. 2c). In total, we identified 1.7 million putative viral genomes, of which 3677 were classified as complete, 15 481 were classified as high-, and 30 484 were classified as medium quality, and were used in subsequent analyses (Supplementary Fig.1). Overall, 18 268 of the 49 642 genomes (36.8%) were identified within

host sequences, indicating a state of lysogeny. Clustering of viral genomes on a 95% similarity level resulted in 18 494 vOTUs (of which 1475 were comprised of genomes from 5 individuals or more; Supplementary Data 1), representing 37.3% of the potential vOTU diversity by Chao1 estimation of species richness. A mean of 223 vOTUs (sd = 69.3) per sample were observed after mapping sequencing reads to vOTU representative sequences (Fig. 2b). Inverse Simpson's diversity index ranged between 2.79 and 245 (mean = 93.5, sd = 43.7). With regards to beta-diversity, the Bray-Curtis dissimilarity index ranged between 0.43 and 1 (mean = 0.84, sd = 0.065; Supplementary Table 1).

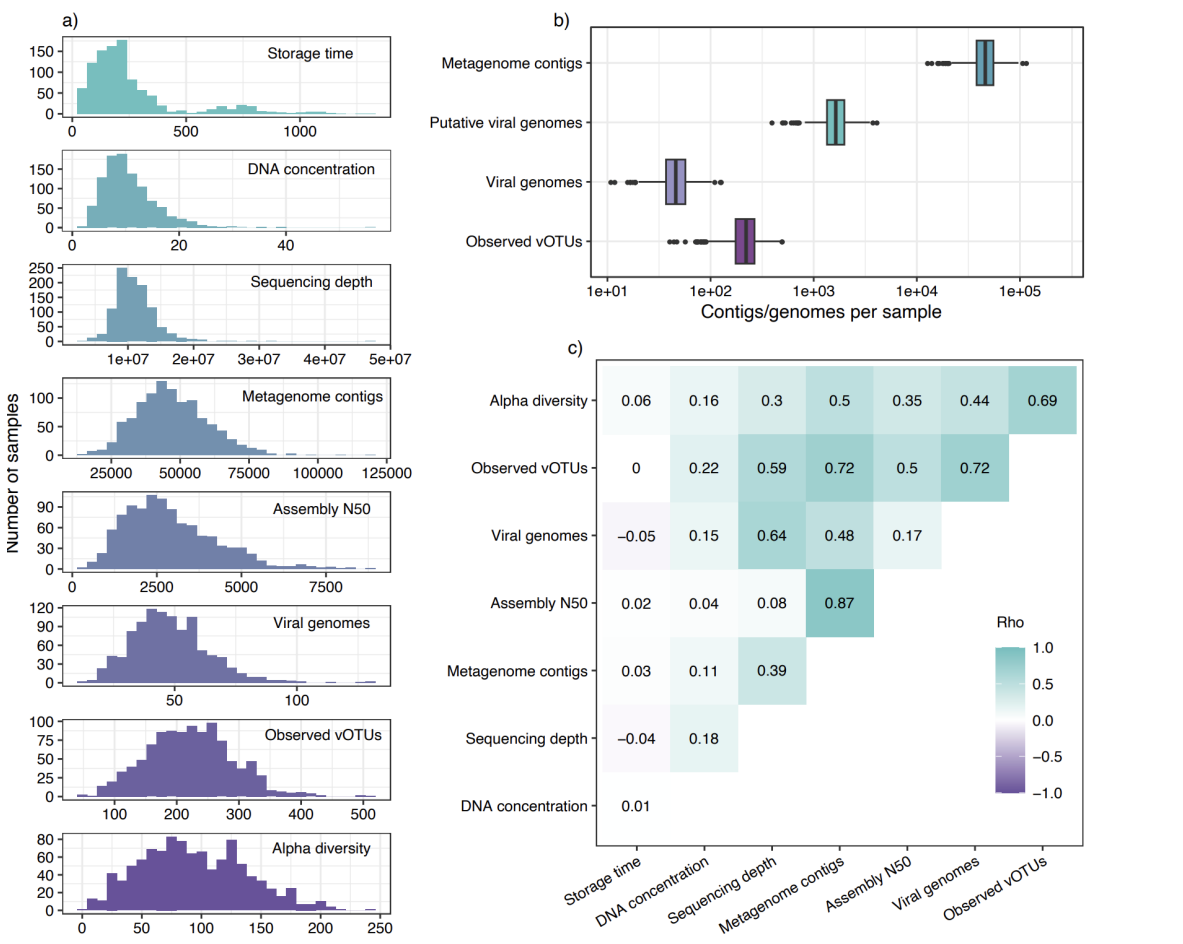

Figure 2: Quality assessment of the virome dataset. a) Histograms of measures by sample including storage time, DNA concentration, number of sequencing reads, number of metagenome contigs, assembly N50, number of viral genomes, vOTUs, and alpha diversity (inverse Simpson index). b) Number of contigs or viral genomes at different stages of the analysis per sample: total number of metagenome contigs, putative viral genomes, filtered genomes (medium quality, high quality and complete) and observed vOTUs after read mapping. c) Pairwise Spearman's rank correlation coefficients (rho) of the measures in a).

vOTU taxonomy and functional potential

Of 18 494 vOTUs, 6036 (32.6%) were assigned taxonomy based on their gene similarity to reference genomes. A majority of these vOTUs (n = 4091, 22.1% of all) were only assigned to a taxonomic order

or class, and were more widely dispersed than family-annotated genomes (Fig.3). The vOTUs that were assigned taxonomic family (1135), represented only 6.1% of all vOTUs. Overall, 19 viral families were identified. The most frequent viral family was *Microviridae* (Fig. 4a), with 528 members. Four families, and 416 vOTUs, of the order *Crassvirales* (*Suoliviridae*, *Intestiviridae*, *Crevaviridae*, and *Steigviridae*) were identified. In addition, the families *Peduoviridae*, *Inoviridae*, and *Winoviridae* were each identified with at least 20 members (Supplementary Table 2). A large fraction of genomes belonging to the class *Caudoviridicetes* belonged to lineages with the former[46] morphology-based classifications *Siphoviridae*, *Myoviridae* and *Podoviridae* (n=2849). The fraction of uncovered vOTU diversity, according to Chao1 estimates, differed by family, with 60% and 74% of *Crevaviridae* and *Winoviridae* respectively, being detected. On the other hand, the detection rates of *Microviridae* and *Inoviridae* were much lower, with 9.9% and 7.3% identified respectively (Supplementary Table 2). Multiple vOTU characteristics differed markedly between viral families, including genome size (Fig. 4b), genome integration (Fig.4c), gene annotation frequency (Fig. 4d), and the rate at which auxiliary metabolic genes were detected (Fig. 4e).

Intestiniviridae, *Suoliviridae*, *Steigviridae*, and *Inoviridae* genomes were almost exclusively identified as unintegrated (Fig. 4c), while genomes of the *Crevaviridae* and *Microviridae* families had a small, but not insignificant fraction of integrated genomes. On the other hand, most genomes of the *Peduoviridae* and *Winoviridae* families were identified in an integrated state.

AMGs were detected in 24.3% of vOTUs, being more commonly detected in *Crassvirales* (67.5%), and less common in *Microviridae* vOTUs (1.1%). AMGs from “Organic nitrogen” and “Miscellaneous” functional groups were detected in 12.8% and 11.7% of vOTUs, respectively, being about five times more prevalent than any other functional group or combinations of these (Supplementary Fig. 2). On a family level, the prevalence of the “Organic nitrogen” group of AMGs was almost absent from vOTUs belonging to *Crassvirales* (0.2%), being largely confined to viruses of the *Peduoviridae* family, and those without a family annotation (Fig. 4e). AMGs of the “Miscellaneous” group (almost exclusively genes related to pyrimidine deoxyribonucleotide synthesis) were detected in a majority (67.1%) of the *Crassvirales* vOTUs, and in particular those belonging to *Steigviridae* (78.8%) and *Intestiniviridae* (88.1%).

Abundance was assessed by mapping reads from all samples to each vOTU. This increased the total number of detected viruses in each sample (mean identified genomes per sample 48; mean observed vOTUs 215). Out of 18 494 vOTUs, 2576 were detected in ≥1% of the population. A mean of 24.4% of viral abundance by sample were attributed to vOTUs with any taxonomic annotation (range 7.9-83.0%; Fig. 4f). *Crassvirales* vOTUs were detected in 70.6% of samples and constituted up to 75.4% of viral abundance (median 0.6%). Overall, *Crassvirales* vOTUs, and especially those of the *Intestiviridae* family, were more abundant when detected, whereas *Microviridae* and *Peduoviridae* were less abundant.

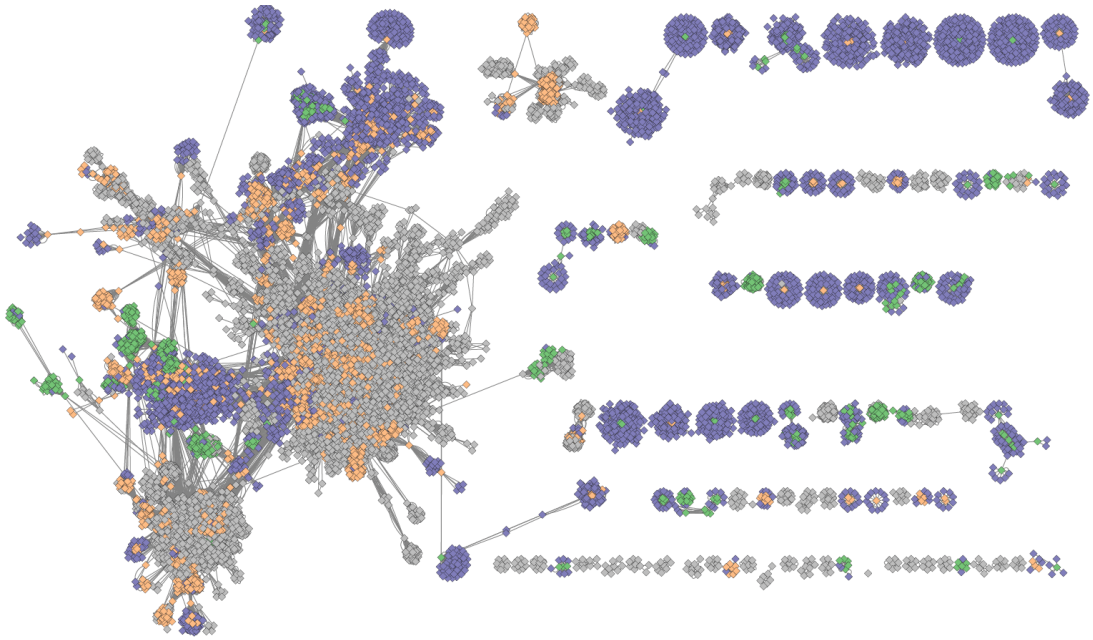

Figure 3: Clustering of the vOTUs based on their gene similarity on a protein level. Green - vOTUs that had taxonomic family annotation, orange - vOTUs that were assigned taxonomic order, but not family, grey - vOTUs with no taxonomic assignment, purple - reference viral genomes. Outlier vOTUs (those with no significant associations) were excluded from visualization.

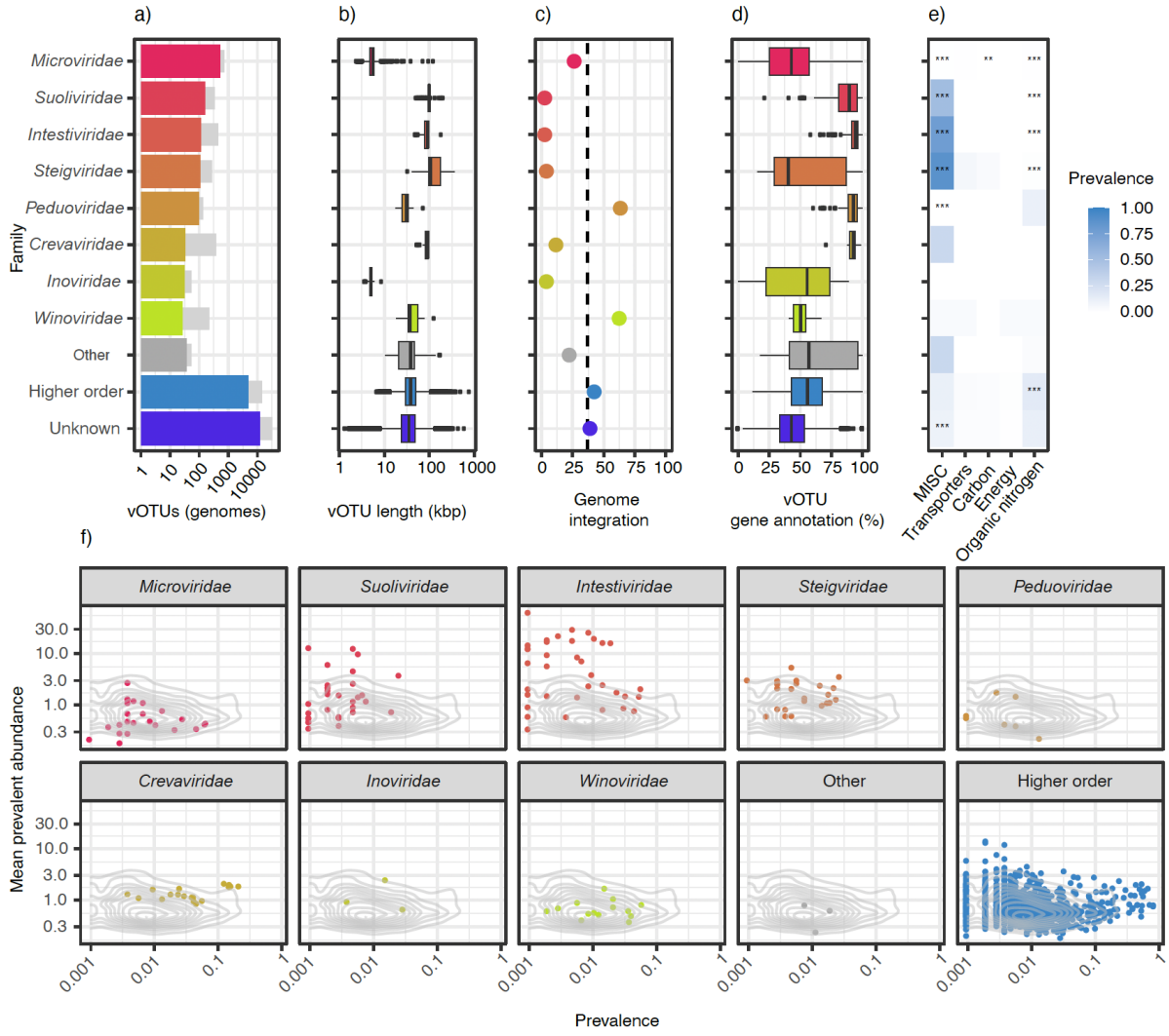

Figure 4: Genome annotation and population distribution. (a) Taxonomic classification of vOTUs at the family level. The vOTUs belonging to families with fewer than 20 representatives are categorized as "other". The "unknown" group constitutes those not clustering with any reference genomes, whereas those clustering with reference genomes annotated at higher levels are labeled "higher order". Light gray bars indicate the total number of genomes (pre-dereplication) according to the taxonomic assignment of their representative vOTUs. b) Genome size distribution for genomes belonging to each taxonomic category. Genomes included in the plot include those that are not classified as "complete"; for stratification by completeness, see Supplementary fig. 1. c) The percentage of viral genomes classified as integrated. The dashed line represents the overall percentage of integrated genomes. d) Percentage of annotated genes per vOTUs according to viral family. e) The fraction of genomes carrying genes annotated with AMGs by AMG category and family. Asterisks indicate significant deviations in AMG category prevalence for one family when compared to the rest (post-hoc Fisher exact test, *p < 0.05, **p < 0.01, ***p<0.001; p-adjustment by Bonferroni). MISC: Miscellaneous; Carbon: Carbon utilization f) Prevalence and mean abundance (if detected) for the vOTUs with at least 2 constituent genomes by taxonomic assignment. The 2D density contour lines indicate the overall distribution of prevalence and abundance for vOTUs (≥ 2 constituent genomes).

The gut virome reflects individual health-related lifestyle, including smoking, physical activity and carbohydrate intake

We assessed differences in virome alpha and beta diversity to determine how the gut virome varied by individuals' diet, lifestyle and demography. Out of 25 selected variables (Supplementary Table 3), we identified 9 significant associations with alpha diversity as measured by the inverse Simpson's index (Fig. 5a). Among these, the largest effect sizes were found for physical activity (positive association), alcohol consumption (positive association), and dietary carbohydrate consumption (negative association). Viral beta diversity was significantly associated with 17/25 variables assessed (Fig. 5b), several being health-related lifestyle factors. Indeed, the strongest association was observed for a composite HLI, with other lifestyle variables being relatively strongly associated, including dietary fiber consumption, physical activity, and smoking, among others. Assessing differential abundance of individual vOTUs, we identified several representative genomes being associated with the same set of variables (Fig. 5c). Here, the highest numbers of differentially abundant vOTUs were found for smoking and physical activity (Fig. 5d). Dietary fiber consumption was also associated with a high number of differentially abundant vOTUs (Fig. 5d, Supplementary Fig.4). Among differentially abundant vOTUs, there was no skew in the frequency of any viral families, nor with the frequency of viruses with a lytic or lysogenic lifestyle (data not shown). On the other hand, we observed a clear over-representation of AMGs across the differentially abundant vOTUs (Supplementary Fig. 3), being especially evident for those related to smoking. Due to the inclusion of participants from a high-risk screening population, there was an over representation of colorectal cancer. To assess whether this might have influenced the observed associations, we performed sensitivity analyses excluding any participants with colorectal cancer and found no overall differences in identified associations (Supplementary Fig. 5a-c).

Overall, 69 vOTUs were related to at least one lifestyle or demographic variable, with 22 being associated with multiple. As an example, one vOTU (CRCbiome_vOTU05693, no taxonomic assignment) was negatively associated with smoking, and positively correlated with physical activity and dietary fiber consumption (Fig. 5d). This vOTU was identified in 62.2% of participants, and was representative of 23 viral genomes, none of which were found to be integrated in a host genome. Gene annotation (44% of predicted genes) identified genes encoding an integrase, an DNA topoisomerase and two methyltransferases (Fig. 5e), indicating a potential capacity of this vOTU to integrate a bacterial host genome. DNA methylase, which is crucial for host defense and epigenetic regulation, was also identified in the CRCbiome_vOTU05693 genome.

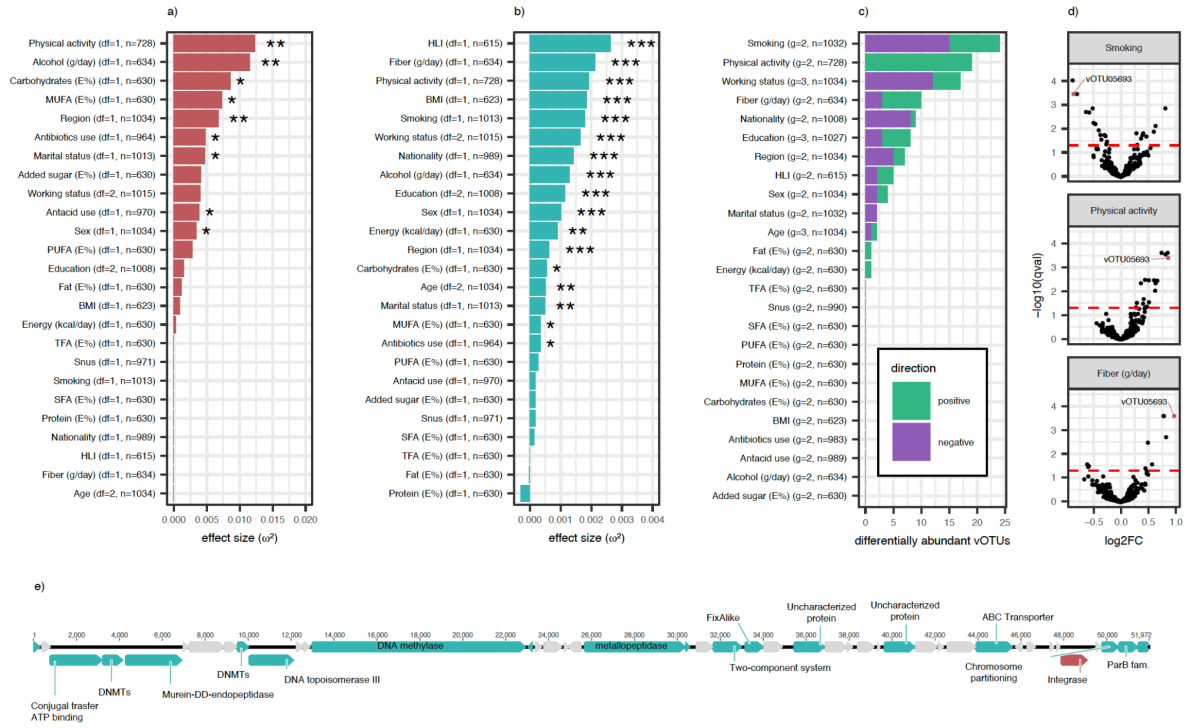

Figure 5: Associations of viral diversity with diet, lifestyle and demographic variables. a) Effect sizes of alpha diversity of vOTU abundance as measured by the inverse Simpson index by ANOVA. b) Effect sizes of associations between vOTU beta diversity (Bray-Curtis index) by PERMANOVA. Effect sizes for alpha and beta diversity are derived using the omega squared measure from ANOVA tests of the association between diversity measures and each variable, with correction for sample sequencing coverage. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$. c) Number of significantly differentially abundant vOTUs identified by MaAsLin2, colored by direction of association. For continuous variables, the top and bottom tertiles were compared. d) Volcano plots showing the relationship between effect size (log2 fold change) and significance level (q-value) for vOTUs for physical activity, smoking and fiber intake, from top to bottom. The red dotted line indicates the significance threshold. MUFA: mono-unsaturated fatty acids, PUFA: poly-unsaturated fatty acids, TFA: trans fatty acids, SFA: short-chain fatty acids, BMI: body mass index, HLI: healthy lifestyle index. e) Genomic map representation of CRCbiome_vOTU05693, associated with smoking, physical activity and dietary fiber intake, with predicted genes with annotations in green, without annotations in gray, and integrase gene annotation highlighted in red.

Discussion

The gut microbiome, and the gut virome in particular, has largely been studied using either fresh stool samples or stool samples preserved in buffers designed for snap-shot stabilization of the microbiome[59]. Here we show that analysis of the gut virome using samples collected in a routine setting and stored in a FIT buffer designed for hemoglobin stabilization is feasible. The reliability of the FIT sampling kits in the analysis of bacteria has repeatedly been demonstrated[24,60,61], but to the best of our knowledge, the present study is the first to demonstrate this for viruses. Use of FIT samples enabled an

in-depth characterization of the viral constituents of the human gut, and allowed us to discern associations between the gut virome and important health-related lifestyle factors, although interpretation of findings remains hampered by the incompleteness of reference databases.

Even though FIT samples are designed to capture as little as 10 mg of fecal matter, only a minor fraction of samples (<1%) failed to produce sequencing data, and viruses were identified in all samples with sufficient data. Stability under storage conditions and DNA quality and quantity are key for the reliability of generated data. Our finding that DNA concentration, sequencing depth and viral diversity were only negligibly affected by sample storage duration lend support to the use of FIT kits as a suitable sampling methodology for virome characterization. FIT sampling is widely employed in population based colorectal cancer screening programs, highlighting the potential for large-scale virome studies across the world.

In this extensive analysis of the gut virome in 1034 Norwegian adults, we identified over 18 000 vOTUs representing more than 49 000 complete, high- or medium quality viral genomes detected across the population. Despite a large sample size for a relatively homogeneous population, our estimates of species richness show that increased sampling would be required to more fully describe the gut virome in this setting. Moreover, due to the exclusive measure of DNA as a source of genetic information, our analyses do not include RNA viruses. Still, the uncovered viral diversity is substantial, and is in line with studies using microbiome-adapted sampling methodology[2,3]. Two thirds of the vOTUs detected in our study were not represented in current state-of-the-art reference databases. Furthermore, only one fifth of those that were represented, were assigned taxonomy at the level of family, clearly demonstrating the lack of data on the human virome. Using the newly ratified taxonomy[44], we found *Microviridae* to be the most commonly assigned viral family among the vOTUs, with most *Microviridae* vOTUs being representative of a small number of genomes. On the other hand, vOTUs annotated as *Crevaviridae*, one of the families belonging to *Crassvirales* order, consisted of significantly larger clusters of genomes, indicating that a larger fraction of *Crevaviridae* genomes were identified when compared to *Microviridae*. This finding of a highly diverse group of *Microviridae* vOTUs is in line with current understanding of this viral family; the high rate of mutations and recombination in their characteristically small genomes not only facilitates rapid evolution and adaptation, but also leads to high intra-family diversity[62].

Along with *Crevaviridae* viruses, other viruses of the *Crassvirales* order displayed lower diversity, and, with the exception of the *Steigviridae* viruses, had a higher fraction of genes annotated. Viruses of the *Steigviridae* family have likely followed an independent evolutionary path from other *Crassvirales* viruses, potentially acquiring novel genes and functions via mechanisms like horizontal gene transfer[63]. Other observed characteristics of the *Crassvirales* viruses such as their size (97-131 kb), almost exclusively lysogenic nature, and high prevalence and abundance, are consistent with other studies[14,64].

We found about a third of viral genomes to be integrated in the genome of its host. Genome integration is a common manifestation of a lysogeny, employed by temperate viruses. Lysogeny is one of two predominant viral life cycles, with the other being the lytic one[65]. The lytic cycle involves viral replication, resulting in host cell destruction and the release of new viruses. In contrast, the lysogenic cycle represents a dormant state, wherein the viral genome is replicated in sync with its host, often being integrated into the host genome, creating a prophage which can be activated to revert to the lytic cycle under certain conditions. Strategies for the study of phage lifecycles include the identification of phages with a potential for transition to a lysogenic state, and direct detection of host genome insertion[66,67]. The former of these is hampered by poor database coverage, and does not provide a measure of actual lysogeny, whereas the latter, which we employed, does provide such a measure, but does not count phages whose lysogenic state occurs in a rolling cycle replicating or plasmid-like state within the host cell. There were clear differences between viral families in their propensity for genome integration, where in contrast to the almost exclusively lytic *Crassvirales* and *Inoviridae* viruses, two viral families, *Peduoviridae* and *Winoviridae*, contained mainly prophages. Interestingly, in a recent study on prophages in infants and adults, *Peduoviridae* was among the most frequently detected, whereas *Winoviridae* phages were not listed[68].

Auxiliary metabolic genes (AMGs) are important for phage modulation of bacterial function[69]. The two most common AMG categories identified in the current population included nitrogen metabolism and nucleotide synthesis (pyrimidine deoxyribonucleotide synthesis, or MISC in Fig. 3e). These AMGs can enhance viral replication efficiency by boosting the bacterial host's pyrimidine synthesis and providing a selective advantage to the virus. This could disrupt the bacterial host's pyrimidine balance, leading to potential cell resource misallocation, nucleotide overproduction, or DNA damage. The small genomes of the *Microviridae* contained few AMGs. In general, when detected, viral genomes tended to contain multiple AMGs per genome. AMGs were common in *Crassvirales* vOTUs, with nucleotide synthesis genes being over-represented and organic nitrogen AMGs being under-represented. Genes involved in metabolism of organic nitrogen were primarily found in the *Peduoviridae* family and within vOTUs that remained unclassified at the family level.

Lifestyle factors have been shown to exhibit significant associations with the bacteria of the gut[70]. However, far less is known for the viral fraction. We conducted a comprehensive analysis of how viral abundance was related to individual diet, lifestyle and demographics factors, measured in broad and generalizable terms. Virome alpha diversity displayed some variation, but not as pronounced as the beta diversity. We found lifestyle factors such as physical activity, dietary fiber and alcohol consumption to have consistent associations with gut virome alpha and beta diversity. Although differences in lifestyle assessment and categorization make direct comparisons difficult, recent studies of various populations have found alcohol intake, as well as diets reflecting a higher intake of fiber to be associated with virome characteristics[3,13,14], while no associations were found for physical activity.

Smoking has been extensively studied for its genetic and epigenetic effects in human cells[71,72]. We found smoking to be associated with beta diversity, in line with some[3], but not all[13] prior reports. Contrary to what has been reported previously[2], we did not find an association between gut virome composition and participant age. However, our results are in line with a recent report showing maintained diversity in subjects of advanced age[73].

Consistent with beta-diversity differences, individual vOTUs were differentially abundant according to subject lifestyle. Differentially abundant vOTUs displayed no propensity towards particular viral clades, nor genome integration state, but we did observe an intriguing over-representation of AMGs, particularly for vOTUs associated with smoking. Notably, we find that several of them were differentially abundant with regard to a number of diet, lifestyle and demographic factors. Moreover, an index capturing multiple aspects of a healthy lifestyle (healthy lifestyle index; HLI) was found to have the largest effect size in relation to gut virome beta diversity. This suggests that several lifestyle factors that affect health may act in concert to shift virome composition. There has been a recent trend in public health research focusing on the overall pattern of lifestyle choices, rather than individual factors[74].

An example illustrative of the challenges and promise of gut virome analyses was our identification of CRCbiome_vOTU05693 as being negatively associated with smoking, and positively associated with physical activity and dietary fiber intake. While being a possibly important indicator of a health associated lifestyle, no taxonomic information was possible to derive from current reference databases. None of the annotated genes were AMGs, but indicated a capacity for host genome integration, host defense, epigenetic gene regulation and maintenance of genome stability[75]. Still, none of its 23 constituent genomes were identified in an integrated state. These observations highlight the need for continued studies and expansion of reference databases for the gut virome, and functional studies of particular viruses.

Collectively, these associations indicate that lifestyle choices may influence the composition and viral make-up of the gut virome. While evidence is limited, recent intervention studies have shown that a short-term change of diet can lead to significant alterations in both the human and mouse gut virome[11,76]. It is likely, though, that alterations in viral abundances are accompanied by, or even precipitated by shifts in abundance of their bacterial hosts.

The main strength of this study includes a large population, which draws on participant recruitment carried out as part of a population-based Norwegian screening trial, inviting all residents of a defined age range and geographic region[27]. Standardized data collection included rich and high-quality data on participant diet and lifestyle. Minimal technical interference in the high quality metagenomes enabled detailed analyses of virome taxonomy, annotation and lifecycle. Comprehensive analyses of alpha and beta diversity, vOTUs differential abundance, and the nuances between them, provide a multi-faceted depiction of the virome. Despite these strengths, there are limitations to consider. The participants had a FIT positive test, meaning that they had traces of blood in their stool samples.

Therefore, the proportion of individuals with premalignant or malignant colorectal cancer lesions is higher than in the general population. Sensitivity analyses excluding participants with a malignancy did not, however, impact the study outcomes.

This study shows that the virome can be reliably profiled using FIT samples, by identifying more than 18000 vOTUs from over 1000 individuals and identifies the virome as being deeply connected to host lifestyle and demography. The associations between the gut virome and subject lifestyle suggests a potential for the gut virome to serve as a source of biomarkers. While microbiome studies have identified gut bacteria as disease biomarkers[77], development of viral biomarkers will require large-scale studies defining sources and measures of gut virome variation.

Acknowledgment

The sequencing was performed at the Sequencing laboratory of Institute for Molecular Medicine Finland (FIMM) Technology Centre, University of Helsinki. We would specifically acknowledge Harri Kangas and Pekka Ellonen for their invaluable advice and help. Jan Inge Nordby and Cecilie Bucher-Johannessen contributed to sample preparation and DNA extraction. We would also like to thank Elina Vinberg and Maja Sigerseth Jakobsen for coordinating the CRCbiome project and Erik Natvig for contributing to data management. Geir Hoff, Thomas de Lange, Øyvind Holme, Kristin Randel and Giske Ursin have contributed with establishing the CRCbiome study and nesting it to the Bowel Cancer Screening in Norway study. The Services for sensitive data (TSD) staff has contributed with timely solving computer issues.

This project would not have been possible without funding from the Norwegian Cancer Society, projects 190179 and 198048, the South Eastern Norway Regional Health Authority projects 2022067 and European Union's Horizon 2020 Research and Innovation program under the Marie Sklodowska-Curie Action Grant agreement No 801133 - Scientia Fellow. The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Author contribution

TBR, PB, TR contributed to the study conception and design. PB, TBR, ASK, EB, VB participated in the data collection. PI, EB, ASK, EA performed the data analysis. TBR, PI, EB, ASK, EA, WMdV interpreted the results. TBR, PI, EB, ASK, EA prepared the manuscript. All authors have read and approved the final manuscript.

Data availability

FASTA files of CRCbiome vOTUs detected in 5 or more individuals, are available at ENA with accession number (data submission in process). Statistics on vOTUs are available as supplementary

Data 1. Due to the sensitive nature of the remaining data derived from human subjects, processing of data and/or biological material from this project must comply with the General Data Protection Regulation (GDPR). Data processing must have approval from the Regional Committee for Medical Research in Norway (REC). Furthermore, the processing needs legal basis according to GDPR Article 6 and 9 and the need for a Data Protection Impact Assessment (DPIA) according to GDPR article 35 must be considered. Requests for data access can be directed to Trine B Rounge, trinro@uio.no.

Code availability

The custom Snakemake pipeline and R scripts used in this study are available at: https://github.com/Rounge-lab/CRCbiome_virome_2023

References

1. Nayfach, S. *et al.* Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. *Nat. Microbiol.* **6**, 960–970 (2021).
2. Gregory, A. C. *et al.* The Gut Virome Database Reveals Age-Dependent Patterns of Virome Diversity in the Human Gut. *Cell Host Microbe* **28**, 724-740.e8 (2020).
3. Nishijima, S. *et al.* Extensive gut virome variation and its associations with host and environmental factors in a population-level cohort. *Nat. Commun.* **13**, 5252 (2022).
4. Shah, S. A. *et al.* Expanding known viral diversity in the healthy infant gut. *Nat. Microbiol.* **8**, 986–998 (2023).
5. Camarillo-Guerrero, L. F., Almeida, A., Rangel-Pineros, G., Finn, R. D. & Lawley, T. D. Massive expansion of human gut bacteriophage diversity. *Cell* **184**, 1098-1109.e9 (2021).
6. Zuppi, M., Hendrickson, H. L., O'Sullivan, J. M. & Vatanen, T. Phages in the Gut Ecosystem. *Front. Cell. Infect. Microbiol.* **11**, (2022).

7. Borodovich, T., Shkoporov, A. N., Ross, R. P. & Hill, C. Phage-mediated horizontal gene transfer and its implications for the human gut microbiome. *Gastroenterol. Rep.* **10**, goac012 (2022).
8. Schroven, K., Aertsen, A. & Lavigne, R. Bacteriophages as drivers of bacterial virulence and their potential for biotechnological exploitation. *FEMS Microbiol. Rev.* **45**, fuaa041 (2021).
9. Montassier, E. *et al.* Probiotics impact the antibiotic resistance gene reservoir along the human GI tract in a person-specific and antibiotic-dependent manner. *Nat. Microbiol.* **6**, 1043–1054 (2021).
10. Federici, S., Nobs, S. P. & Elinav, E. Phages and their potential to modulate the microbiome and immunity. *Cell. Mol. Immunol.* **18**, 889–904 (2021).
11. Minot, S. *et al.* The human gut virome: inter-individual variation and dynamic response to diet. *Genome Res.* **21**, 1616–1625 (2011).
12. Shkoporov, A. N. *et al.* The Human Gut Virome Is Highly Diverse, Stable, and Individual Specific. *Cell Host Microbe* **26**, 527-541.e5 (2019).
13. Zuo, T. *et al.* Human-Gut-DNA Virome Variations across Geography, Ethnicity, and Urbanization. *Cell Host Microbe* **28**, 741-751.e4 (2020).
14. Gulyaeva, A. *et al.* Discovery, diversity, and functional associations of crAss-like phages in human gut metagenomes from four Dutch cohorts. *Cell Rep.* **38**, 110204 (2022).
15. Yang, K. *et al.* Alterations in the Gut Virome in Obesity and Type 2 Diabetes Mellitus. *Gastroenterology* **161**, 1257-1269.e13 (2021).
16. Clooney, A. G. *et al.* Whole-Virome Analysis Sheds Light on Viral Dark Matter in Inflammatory Bowel Disease. *Cell Host Microbe* **26**, 764-778.e5 (2019).

17. Castellarin, M. *et al.* Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma. *Genome Res.* **22**, 299–306 (2012).
18. Haghi, F., Goli, E., Mirzaei, B. & Zeighami, H. The association between fecal enterotoxigenic B. fragilis with colorectal cancer. *BMC Cancer* **19**, 879 (2019).
19. Bucher-Johannessen, C. *et al.* Long-term follow-up of colorectal cancer screening attendees identifies differences in Phascolarctobacterium spp. using 16S rRNA and metagenome sequencing. *Front. Oncol.* **13**, (2023).
20. Scott, A. J. *et al.* International Cancer Microbiome Consortium consensus statement on the role of the human microbiome in carcinogenesis. *Gut* **68**, 1624–1632 (2019).
21. Hannigan, G. D., Duhaime, M. B., Ruffin, M. T. th, Koumpouras, C. C. & Schloss, P. D. Diagnostic Potential and Interactive Dynamics of the Colorectal Cancer Virome. *mBio* **9**, (2018).
22. Navarro, M., Nicolas, A., Ferrandez, A. & Lanas, A. Colorectal cancer population screening programs worldwide in 2016: An update. *World J. Gastroenterol.* **23**, 3632 (2017).
23. Allison, J. E., Fraser, C. G., Halloran, S. P. & Young, G. P. Population Screening for Colorectal Cancer Means Getting FIT: The Past, Present, and Future of Colorectal Cancer Screening Using the Fecal Immunochemical Test for Hemoglobin (FIT). *Gut Liver* **8**, 117–130 (2014).
24. Rounge, T. B. *et al.* Evaluating gut microbiota profiles from archived fecal samples. *BMC Gastroenterol.* **18**, 171–171 (2018).
25. Krigul, K. L., Aasmets, O., Lüll, K., Org, T. & Org, E. Using fecal immunochemical tubes for the analysis of the gut microbiome has the potential to improve colorectal cancer screening. *Sci. Rep.* **11**, 19603 (2021).

26. Birkeland, E. *et al.* Profiling small RNAs in CRC screening samples such as the widely used fecal immunochemical test, is it possible? 2023.05.03.23289251 Preprint at https://doi.org/10.1101/2023.05.03.23289251 (2023).
27. Kværner, A. S. *et al.* The CRCbiome study: a large prospective cohort study examining the role of lifestyle and the gut microbiome in colorectal cancer screening participants. *BMC Cancer* **21**, 930 (2021).
28. Brunvoll, S. H. *et al.* Validation of repeated self-reported n-3 PUFA intake using serum phospholipid fatty acids as a biomarker in breast cancer patients during treatment. *Nutr. J.* **17**, 94 (2018).
29. Carlsen, M. H. *et al.* Evaluation of energy and dietary intake estimates from a food frequency questionnaire using independent energy expenditure measurement and weighed food records. *Nutr. J.* **9**, 37 (2010).
30. Andersen, L. F. *et al.* Evaluation of three dietary assessment methods and serum biomarkers as measures of fruit and vegetable intake, using the method of triads. *Br. J. Nutr.* **93**, 519–527 (2005).
31. Matvaretabellen. https://www.matvaretabellen.no/.
32. Shams-White, M. M. *et al.* Operationalizing the 2018 World Cancer Research Fund/American Institute for Cancer Research (WCRF/AICR) Cancer Prevention Recommendations: A Standardized Scoring System. *Nutrients* **11**, 1572 (2019).
33. Shams-White, M. M. *et al.* Further Guidance in Implementing the Standardized 2018 World Cancer Research Fund/American Institute for Cancer Research (WCRF/AICR) Score. *Cancer Epidemiol. Biomarkers Prev.* **29**, 889–894 (2020).
34. Helsedirektoratet (The Norwegian Directorate of Health). *Anbefalinger om kosthold, ernæring og fysisk aktivitet (Recommendations for diet. nutrition and physical activity).*

https://www.helsedirektoratet.no/rapporter/anbefalinger-om-kosthold-ernaering-og-fysisk-aktivitet (2014).

35. *Global Recommendations on Physical Activity for Health*. (World Health Organization, 2010).
36. Piercy, K. L. *et al.* The Physical Activity Guidelines for Americans. *JAMA* **320**, 2020–2028 (2018).
37. Kværner, A. S. *et al.* Associations of the 2018 World Cancer Research Fund/American Institute of Cancer Research (WCRF/AICR) cancer prevention recommendations with stages of colorectal carcinogenesis. *Cancer Med.* **12**, 14806–14819 (2023).
38. Kieser, S., Brown, J., Zdobnov, E. M., Trajkovski, M. & McCue, L. A. ATLAS: a Snakemake workflow for assembly, annotation, and genomic binning of metagenome sequence data. *BMC Bioinformatics* **21**, 257–257 (2020).
39. Bushnell, B. BBMap: BBMap short read aligner, and other bioinformatic tools. *SourceForge* https://sourceforge.net/projects/bbmap/ (2022).
40. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res* **27**, 824–834 (2017).
41. Guo, J. *et al.* VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome* **9**, 37–37 (2021).
42. Nayfach, S. *et al.* CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat Biotechnol* **39**, 578–585 (2021).
43. Woodcroft, B. J. Galah - More scalable dereplication for metagenome assembled genomes. (2023).
44. Bin Jang, H. *et al.* Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat. Biotechnol.* **37**, 632–639 (2019).

45. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).

46. Cook, R. *et al.* INfrastructure for a PHAge REference Database: Identification of Large-Scale Biases in the Current Collection of Cultured Phage Genomes. *PHAGE* **2**, 214–223 (2021).

47. Pandolfo, M., Telatin, A., Lazzari, G., Adriaenssens, E. M. & Vitulo, N. MetaPhage: an Automated Pipeline for Analyzing, Annotating, and Classifying Bacteriophages in Metagenomics Sequencing Data. *mSystems* **7**, e00741-22 (2022).

48. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).

49. Shaffer, M. *et al.* DRAM for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic Acids Res.* **48**, 8883–8900 (2020).

50. Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).

51. Thannesberger, J. *et al.* Viruses comprise an extensive pool of mobile genetic elements in eukaryote cell cultures and human clinical samples. *FASEB J.* **31**, 1987–2000 (2017).

52. Aramaki, T. *et al.* KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinforma. Oxf. Engl.* **36**, 2251–2252 (2020).

53. Wang, Y. *et al.* A crowdsourcing open platform for literature curation in UniProt. *PLOS Biol.* **19**, e3001464 (2021).

54. Zhang, H. *et al.* dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res.* **46**, W95–W101 (2018).

55. Brister, J. R., Ako-adjei, D., Bao, Y. & Blinkova, O. NCBI Viral Genomes Resource. *Nucleic Acids Res.* **43**, D571–D577 (2015).

56. vegan: Community Ecology Package. R package version 2.5-7. (2020).

57. Mallick, H. *et al.* Multivariable association discovery in population-scale meta-omics studies. *PLOS Comput. Biol.* **17**, e1009442 (2021).

58. Olejnik, S. & Algina, J. Generalized Eta and Omega Squared Statistics: Measures of Effect Size for Some Common Research Designs. *Psychol. Methods* **8**, 434–447 (2003).

59. Tang, Q. *et al.* Current Sampling Methods for Gut Microbiota: A Call for More Precise Devices. *Front. Cell. Infect. Microbiol.* **10**, (2020).

60. Gudra, D. *et al.* A widely used sampling device in colorectal cancer screening programmes allows for large-scale microbiome studies. *Gut* **68**, 1723–1725 (2019).

61. Masi, A. C. *et al.* Using faecal immunochemical test (FIT) undertaken in a national screening programme for large-scale gut microbiota analysis. *Gut* **70**, 429–431 (2021).

62. Minot, S. *et al.* Rapid evolution of the human gut virome. *Proc. Natl. Acad. Sci.* **110**, 12450–12455 (2013).

63. Ramos-Barbero, M. D. *et al.* Characterization of crAss-like phage isolates highlights Crassvirales genetic heterogeneity and worldwide distribution. *Nat. Commun.* **14**, 4295 (2023).

64. Yutin, N. *et al.* Discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut. *Nat. Microbiol.* **3**, 38–46 (2018).

65. Zhang, M., Zhang, T., Yu, M., Chen, Y.-L. & Jin, M. The Life Cycle Transitions of Temperate Phages: Regulating Factors and Potential Ecological Implications. *Viruses* **14**, 1904 (2022).

66. Arnau, V. *et al.* Inference of the Life Cycle of Environmental Phages from Genomic Signature Distances to Their Hosts. *Viruses* **15**, 1196 (2023).

67. Sutcliffe, S. G., Reyes, A. & Maurice, C. F. Bacteriophages playing nice: Lysogenic bacteriophage replication stable in the human gut microbiota. *iScience* **26**, 106007 (2023).

68. Dikareva, E. *et al.* An extended catalogue of integrated prophages in the infant and adult fecal microbiome shows high prevalence of lysogeny.

69. Luo, X.-Q. *et al.* Viral community-wide auxiliary metabolic genes differ by lifestyles, habitats, and hosts. *Microbiome* **10**, 190 (2022).

70. Asnicar, F. *et al.* Microbiome connections with host metabolism and habitual diet from 1,098 deeply phenotyped individuals. *Nat Med* (2021) doi:10.1038/s41591-020-01183-8.

71. DeMarini, D. M. Genotoxicity of tobacco smoke and tobacco smoke condensate: a review. *Mutat. Res. Mutat. Res.* **567**, 447–474 (2004).

72. Joehanes, R. *et al.* Epigenetic Signatures of Cigarette Smoking. *Circ. Cardiovasc. Genet.* **9**, 436–447 (2016).

73. Johansen, J. *et al.* Centenarians have a diverse gut virome with the potential to modulate metabolism and promote healthy lifespan. *Nat. Microbiol.* **8**, 1064–1078 (2023).

74. World Cancer Research Fund/American Institute for Cancer Research. *Diet, nutrition, physical activity and cancer: A global perspective*. dietandcancerreport.org (2018).

75. Murphy, J., Mahony, J., Ainsworth, S., Nauta, A. & van Sinderen, D. Bacteriophage Orphan DNA Methyltransferases: Insights from Their Bacterial Origin, Function, and Occurrence. *Appl. Environ. Microbiol.* **79**, 7547–7555 (2013).

76. Schulfer, A. *et al.* Fecal Viral Community Responses to High-Fat Diet in Mice. *mSphere* **5**, 10.1128/msphere.00833-19 (2020).

77. Xiao, L., Zhang, F. & Zhao, F. Large-scale microbiome data integration enables robust biomarker identification. *Nat. Comput. Sci.* **2**, 307–316 (2022).