Identification of Phage Genomes in the Gastrointestinal Tracts of Patients with Asthma

Introduction

The identification of novel virus sequences that have come with high-throughput sequencing has led to a boom in new species classifications and reshuffling of viral taxonomy. While historically based in phenotypic analysis like viral host species, genome type (circular or linear, DNA or RNA, single-stranded or double-stranded) and viral morphology, like capsid shape or the presence of a tail, modern viral taxonomy is primarily informed by genomic analysis (Dion *et al*., 2020). The effect on viral taxonomy from this sequencing boom is evident when comparing the 1999 report published by the International Committee on Taxonomy of Viruses (ICTV) which organized 1551 species into 3 orders, 64 families and 234 genera to the 2022 ICTV report which outlined 72 orders, 264 families, and 2818 genera for a total of 11273 species (Walker *et al.,* 2022). Shotgun metagenomic sequencing allows all organisms to be sequenced from a complex sample meaning virus sequences can be collected from a diversity of environments and public access to this data is fueling this sequencing boom.

Taxonomy based in genomic analysis is not the only advantage of recent accessibility and abundance of metagenomic datasets that are publicly available. There are diagnostic opportunities in identifying the viral species present in patients and marking their relative abundance to create a viral profile of disease. For example, Ma et al. (2018) demonstrated 7 novel bacteriophages (phages) shown to be unique to Type 2 Diabetes. There are also treatment prospects that come with better understanding of viral communities in patients as demonstrated by the Faecal Microbiota Transplantation (FMT) allowing bacteriophage transfer for treatment in *Clostridium difficile* infection (Zuo *et al*., 2018). Nonetheless, investigation of the

virome remains challenging because of diversity and evolution in viral genomes (Dion *et al*., 2020) and lack of marker genes like the 16S rRNA used in prokaryotic sequence analysis.

The unknowns of the gut virome have been deemed the “viral dark matter” because faecal samples contain a large portion of viral DNA of unknown origin, accounting for 85-99% of sequences observed in some samples (Aggarwala *et al*., 2017). One investigation showed that the families of viruses that are present in the human gut vary among individuals and remain consistent over several years but within those families there are fluctuations in the dominant populations at the genus and species levels in that time (Garmaeva *et al*., 2021).

For healthy individuals, one of the most abundant genera of bacteria in the intestinal tract are *Bacteroides* (Rinninella *et al.,* 2019). *Bacteroides* is a common host of the most abundant phages in human faecal metagenomic samples called crAssphages, discovered by *de novo* assembly and found in 1.68% of reads from publicly available metagenomic faecal samples at the time (Dutilh *et al.,* 2014). High-throughput sequencing provides an abundance of publicly available metagenomic data that possibly contain novel viral genomes that are ripe for differential and phylogenetic analysis. In addition to simply sequencing previously unknown genomes, there are advantages to modern methodology in sequencing, for example, paired-end reads allow genome assembly without a reference genome conferring improved downstream phylogenetic analysis (Ghurye & Pop, 2019, Prjibelski *et al*., 2014). Further elucidating unknown virus sequences is critical to understanding how the gut virome impacts health and disease.

Several disease states have been associated with their accompanying gut virome and microbiome profiles. The ‘gut-lung axis’ is one such example, where function and health of the respiratory system are linked to the bacterial and viral populations living in the intestines. This

link is supported by respiratory diseases such as asthma and chronic obstructive pulmonary disease (COPD) that have gastrointestinal (GI) co-morbidities like irritable bowel syndrome (IBS) and inflammatory bowel disease (IBD) (Budden et al., 2017). Asthma has a well-known association with the viral and bacterial profiles of the gut particularly in the less stable guts of young children where asthma's impacts are greatest (Frati *et al.*, 2018).

The objective of this study is to use *de novo* genome assembly to reconstruct and describe the phage genomes present in faecal samples collected from patients with moderate-severe asthma, primarily attempting to identify novel phage genomes, with the aim of providing insight into the gut microbiome interactions that are associated with asthma.

Methods

Data Sources and Pre-processing

The data for this investigation was taken from a study by Wilson et al. (2023). Samples were first collected in St. Louis, MO, US in 2015 as part of the Microbiome and Asthma Research Study (MARS). Fecal samples were collected from children and adults aged 6-49 with physician-diagnosed moderate to severe asthma. After extracting crude DNA, samples had DNA purified by a QIAGEN PCR Clean up kit. These samples did not undergo enrichment of the host, bacterial or viral DNA. Following successful library quality control, all samples were tagged with a unique identifying adapter to allow all samples to be sequenced together on an Illumina NovaSeq 6000 platform. High-throughput sequencing yielded paired-end reads for an average of 3.4 Giga-base-pairs per sample. Demultiplexing using unique adapters for each sample produced reads that were published to Sequence Read Archive (SRA) with the accession ERP144764, which contained FASTQ files with Phred quality scores for each base call. To limit computational complexity and space demands, the five smallest FASTQ files of samples from

asthmatic patients were used for this investigation. These reads were downloaded from SRA database using default parameters with the SRAToolkit (v3.0.3).

Initial quality was assessed using FASTQC (v3) (Andrews, 2010) with default parameters and MultiQC (v1.13) (Ewels *et al.,* 2016) with default parameters to visualize FASTQC reports. Using fastp (v0.23.2) (Chen *et al.,* 2018), we performed quality trimming, low-complexity filtering, and adapter removal. Reads were processed with fastp using the "--detect_adapter_for_pe" flag to detect adapters automatically and using the default parameters except for the "-q 20" parameter to limit quality trimming to below Phred quality score of 20 and the "-n 0" parameter to drop reads with any N base calls instead of the 5 N's allowed by default. Bowtie2 (v2.5.1) (Langmead & Salzberg, 2012) aligned cleaned reads in 'local' mode to human reference genome (GRCh38/hg38) using the "-k 1" parameter to limit alignment to 1 distinct alignment only.

Python (v3.8.10) was used throughout the project to process FASTA sequences with the help of the biopython library (v1.81) (Cock *et al.,* 2009). The R programming environment (v4.2.0) was also used to process, summarize, and visualize results primarily with the help of the tidyverse (v1.3.2) packages.

Assembly and Analysis

De novo assembly of contiguous DNA (contigs) assembled by building de Bruijn graphs using k-mer lengths of 21,33,55 was carried out with default parameters using metaSPAdes (v3.15.5) (Nurk *et al*., 2017), a metagenomic modification of SPAdes. Assembly was evaluated using metaQUAST (v5.2.0) (Mikheenko *et al*., 2016), a metagenomic modification of QUAST. We evaluated *de novo* assembly primarily through N50 values given there was not a sole

reference genome that could be used for assembly evaluations like NG50 or misassembly. We also used a custom script in Python (v3.8.10) to generate NX0 values to visualize length distributions (Figure 1).

Collecting contigs that were >10 kbp, we used VirFinder (v1.1) (Ren *et al.,* 2017), a deep-learning model trained to identify contigs of viral origin, particularly because it is expected to perform better than VirSorter in the context of identifying novel viral sequences (Ren *et al.,* 2017). Remaining contigs were used as query sequences with BLAST+ (v2.9.0) against the NCBI reference viruses representative genome database which contains 15,226 viral sequences. BLASTN options included "-evalue 0.001" to limit reported hits to those with an expected value less than 0.001. Of these candidate contigs, analysis only proceeded with those that did not map to any representative viral sequence. To attempt to identify novel phage genomes using these contigs, BLAST results were cross-referenced with metaQUAST reports and contigs that had > 50% alignment to the default SILVA 16S rRNA database used by metaQUAST were filtered out from candidate contigs.

Open Reading Frame (ORF) prediction on the remaining candidate contigs was carried out using Prodigal (v2.6.3) (Hyatt *et al.,* 2010) with default settings. Predicted amino acid sequences were used as query sequences in BLASTP alignment against the NCBI non-redundant protein database. Candidate (partial) genomes were used as a reference in alignment with Bowtie2 (v2.5.1) to map reads from other samples in order to determine if this candidate partial genome could be detected in other samples.

Results

FASTQ files of all samples obtained from SRA contained more than 8 million reads per sample. When initially analyzed to begin quality control and preprocessing, the MultiQC report showed a lack of over-represented sequences, lack of N-base calls and all sequences passed all FASTQC quality checks (except for Per Base Sequence Content which could be explained by PCR amplification using random hexamers). In addition to confirming that quality control and preprocessing had occurred prior, an assessment with fastp showed that essentially no adapter sequences or demultiplexing sequences that required trimming were present in these reads, with a maximum for all samples of 0.21% of reads labelled as requiring trimming.

Alignment to human genome GRCh38 using Bowtie2 yielded minimal number of alignments (rounded to 0.00% of reads for all samples) indicating that beyond pre-processing, the reads that aligned to the human genome were also removed before uploading reads to SRA. As a result, the minimal trimming and human host-filtering that we observed was ignored and FASTQ files directly from SRA were used for genome assembly that followed.

De novo assembly yielded 258K, 339K, 303K, 359K, and 366K contigs for samples 1 – 5, respectively. Figure 1 shows the length distribution (NX0 values) of contigs for each sample. Considering the smallest genome size of known tailed phages is ~11 kbp (Hatfull & Hendrix, 2011) and assembly of each sample yielded >250 000 contigs, a minimum length threshold of 10 kbp was used to limit contigs for further analysis. This length threshold was also used by Camarillo-Guerrero et al. (2021) in the development of the Gut Phage Database.

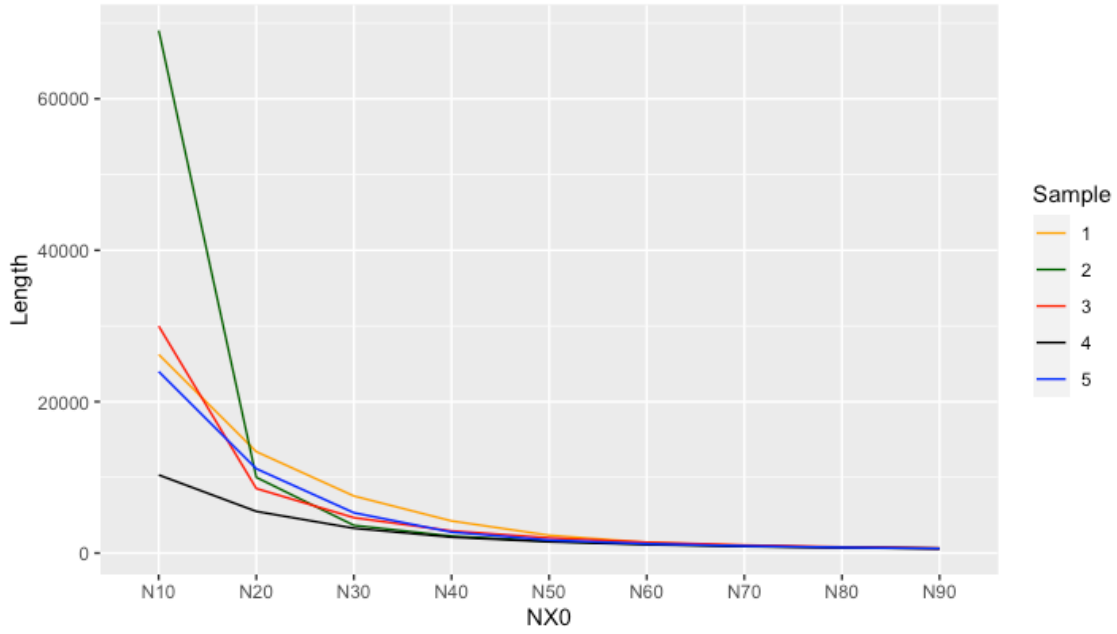

Figure 1: The length (base-pair) distribution of assembled contigs for each of the 5 asthmatic samples with exact N50 values of 2348 for sample 1, 1564 for sample 2, 2008 for sample 3, 1486 for sample 4 and 1719 for sample 5.

Of the 3817 contigs assembled above 10 kbp, VirFinder identified 39 contigs that were likely to have viral origin (score >0.9 and p-value <0.05), moving forward these are referred to as “viral contigs”. All 39 viral contigs were used as query sequences for BLASTN alignments and all but 8 contigs aligned to viral representative genomes with e-value <0.001. BLASTN of viral contigs resulted in alignments to several crAssphage representative genomes with expected values rounded to 0.0 and identities <90% (including Uncultured phage cr99_1 (NC_062771.1), CrAssphage cr11_1 (NC_055876.1), Uncultured phage cr9_1 (NC_062776.1), Uncultured phage cr123_1 (NC_062766.1), CrAssphage cr110_1 (NC_055879.1), CrAssphage sp. isolate ctbg_1 (NC_055760.1) and several others). In addition, there were reported alignments to Faecalibacterium phage FP_Toutatis (NC_047915), Bacillus phage vB_BboS-125 (NC_048735), Clostridium phage c-st (NC_007581), and Streptococcus prophage EJ-1 (NC_005294) all with expected values < 7e-175 and identities >90%.

From the 8 viral contigs that did not align to any representative viral sequences, 1 was removed after cross-referencing with metaQUAST results which indicated it had been misassembled. ORF prediction of the remaining 7 candidate contigs yielded a total of 105 predicted genes with 16 of those being >500 residues in length, used as a length threshold to limit BLAST computational demands. BLASTP alignment against the NCBI non-redundant protein database using these 16 predicted protein sequences identified 1 predicted gene that aligned to 9 sequences in the database coding for Phage Terminase Large Subunit (LSU) with identities >95% and expected values rounded to 0.0. The contig that gave rise to this ORF originated in sample 5 with a coverage of 6 and length of 10103 bp. This contig was used as a reference to map reads from samples 1-4, where 31 reads successfully mapped to this candidate contig from sample 4 only.

Discussion

Limitations

The lack of access to raw reads before pre-processing likely biased the assembly because of the pre-processing performed by Wilson et al. (2023). The choice of aligner for human genome filtering influences downstream assembly results (Islam *et al.,* 2021) as does choices like quality control and adapter trimming. Performing sensitivity analysis on assembly and preprocessing using a variety of preprocessing options could have been possible had the raw reads been available for download. Subtle differences also exist in the genome assembly methodology by Wilson et al. (2023), namely the use of a k-mer length of 77 in addition to the default k-mer lengths of 21, 33, 55 used here. This change is recommended for long reads but only if coverage was high (X50+), however, in the context of novel phage sequence assembly, the coverage is not likely to be as high as when identifying *ermF* sequences as was done by Wilson et al. (2023). The coverage was below 50 for all viral contigs assembled in this study.

The evaluation of *de novo* assembly by metaQUAST was not as informative as it could have been had comparisons to viral reference genomes been possible. metaQUAST reported being unable to download genomes from the NCBI database of provided reference viruses, primarily crAssphages. These sequences were curated from initial BLASTN results from viral contigs as query sequences against the viruses representative genome database. Time constrictions prevented further investigation and analysis proceeded with minimal assembly evaluation using reference sequences from the default SILVA 16S rRNA database. Nonetheless, a contig was eliminated as a candidate contig because misassembly was indicated by metaQUAST. Though metaQUAST was able to report basic evaluation stats like N50 and some misassembly, additional evaluation metrics would have allowed more thorough identification of misassembly, reducing the chance of false positives from cross-assembled contigs. This issue is exacerbated when working with phage genomes because of the gene transfer and rapid evolution that is seen in phage genomes (Dion *et al.,* 2020). Identifying misassembly and deeper evaluation of metagenomic assembly without a reference is now possible with recently developed tools like metaMIC (Lai *et al.,* 2022) and could be explored in the future.

In the pursuit of viral contigs, VirFinder reported a score with p-values but since it uses multiple comparisons throughout a bacterial host and viral genome database, multiple test correction is required to determine adjusted p-values. However, no multiple test correction was performed for this investigation because the p-values yielded from VirFinder produced an error in the attempt to use the built-in qvalue() function. Without being able to determine the number of sequences used in the database for comparison, the use of other multiple test correction

methods like Bonferroni correction was not possible. Without multiple test correction the report of p-value lacks an indication of expected false discovery rate.

Previously Described Phages

Given the gastrointestinal tract is home to an abundance of bacteria, the phages that infect the most common of these bacteria would be expected to be found in genomic assembly of human faecal metagenomic samples and this investigation was no exception. All known phages found here by *de novo* assembly were tailed bacteriophages, of the class *Caudoviricetes*. In healthy intestines of adults, *Bacteroides* and *Faecalibacterium* are two of the most abundant genera from this class (Rinninella *et al.,* 2019). We demonstrated the presence of several crAssphages, which have *Bacteroides* hosts, and Faecalibacterium phages, which have *Faecalibacterium* hosts and their presence is consistent with several other metagenomic studies of human faeces (Bassi *et al.,* 2022, Dutilh *et al.,* 2014, Shkoporov & Hill, 2019).

Though not as numerous, the intestinal presence of other phage genomes assembled in this study, namely Bacillus phage and Clostridium phage, present interesting opportunities for future investigation. Bacillus phages belong to the genus *Andromedavirus* and infect *Baccilus pumilus* group species of bacteria, some of which have been shown to cause food poisoning (From *et al.,* 2007) and can produce spores that survive in extreme conditions, making them more difficult to kill than colonies of other species. Recently, two novel species of the genus *Andromedavirus* have been sequenced from soil and are suspected to possess polysaccharide depolymerases that confer anti-bacterial activity against *Baccilus pumilus* (Skorynina *et al.,* 2022). The further investigation of Bacillus phages found in faecal metagenomic samples could unlock additional anti-bacterial potential. Another phage sequence assembled in this investigation that is more relevant to asthma is that of Clostridium phage – a phage that infects

Clostridium. This genus of bacteria is diverse in its effects in the human intestines – some species such as *Clostridium difficile* prove to be difficult infections to eliminate with costly consequences for the human host (Czepiel *et al.,* 2019). Interestingly, several species have been shown to be enriched in adults with asthma including *Clostridium bolteae, Clostridium ramosum,* and *Clostridium spiroforme* (Wang *et al.,* 2018) which could explain the presence of Clostridium phage observed here. The differential analysis as it relates to asthma and the abundance of these phages that infect *Clostridium* remains a future opportunity for study.

Novel Phages

This investigation yielded several contigs that showed promise as potentially novel phage sequences. Several passed the thresholds used by Camarillo-Guerrero et al. (2021) to be characterized as viral sequences and did not align well to known representative genomes in the NCBI database. However, query to more appropriate and expansive recently developed databases such as the Gut Phage Database (Camarillo-Guerrero *et al.,* 2021) or the IMG/VR database (Camargo *et al.,* 2023) could provide more insight into whether these sequences have been observed previously. The computational demand of querying these databases remained a barrier for this investigation.

As the contig that was marked as the candidate genome is about 10 kpb, it is likely not a full phage genome sequence, with average genome sizes around 50 kbp (Al-Shayeb *et al.,* 2020). Mapping reads from other samples could provide additional reads with which to cluster and reconstruct the full novel genome. Given only 31 reads from sample 4 mapped to this candidate (partial) genome and the small sample size used in this investigation, it is inconclusive whether this implies this phage is somewhat unique to this individual or can be found in the gut more broadly. It would be worthwhile to continue to attempt to map reads from

other samples within the same study and other faecal metagenomic samples available publicly to this candidate genome to attempt to discern if this sequence can be found elsewhere and to what degree.

The alignment of a predicted protein sequence to other Phage Terminase LSU sequences is also worth pursuing because it is found in many phage species and can act as a common gene for future protein-based phylogenetic analysis (Feiss & Rao, 2012). Additional challenges related to the genome mosaicism, gene transfer and rapid evolution of phages (Dion *et al.,* 2020) make genome assembly and phylogenetics particularly difficult. The opportunity in characterizing phageomes associated with disease remains a possible avenue for diagnostic and treatment breakthroughs.

References

Aggarwala, V., Liang, G., & Bushman, F. D. (2017). Viral communities of the human gut: metagenomic analysis of composition and dynamics. *Mobile DNA*, *8*, 12. https://doi.org/10.1186/s13100-017-0095-y

Al-Shayeb, B., Sachdeva, R., Chen, L. X., Ward, F., Munk, P., Devoto, A., Castelle, C. J., Olm, M. R., Bouma-Gregson, K., Amano, Y., He, C., Méheust, R., Brooks, B., Thomas, A., Lavy, A., Matheus-Carnevali, P., Sun, C., Goltsman, D. S. A., Borton, M. A., Sharrar, A., … Banfield, J. F. (2020). Clades of huge phages from across Earth's ecosystems. *Nature*, *578*(7795), 425–431. https://doi.org/10.1038/s41586-020-2007-4

Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. *Available online at*: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Bassi, C., Guerriero, P., Pierantoni, M., Callegari, E., & Sabbioni, S. (2022). Novel Virus Identification through Metagenomics: A Systematic Review. *Life (Basel, Switzerland)*, *12*(12), 2048. https://doi.org/10.3390/life12122048

Budden, K. F., Gellatly, S. L., Wood, D. L., Cooper, M. A., Morrison, M., Hugenholtz, P., & Hansbro, P. M. (2017). Emerging pathogenic links between microbiota and the gut-lung axis. *Nature reviews. Microbiology*, *15*(1), 55–63. https://doi.org/10.1038/nrmicro.2016.142

Camargo, A. P., Nayfach, S., Chen, I. A., Palaniappan, K., Ratner, A., Chu, K., Ritter, S. J., Reddy, T. B. K., Mukherjee, S., Schulz, F., Call, L., Neches, R. Y., Woyke, T., Ivanova, N. N., Eloe-Fadrosh, E. A., Kyrpides, N. C., & Roux, S. (2023). IMG/VR v4: an expanded database of uncultivated virus genomes within a framework of extensive functional, taxonomic, and ecological metadata. *Nucleic acids research*, *51*(D1), D733–D743. https://doi.org/10.1093/nar/gkac1037

Camarillo-Guerrero, L. F., Almeida, A., Rangel-Pineros, G., Finn, R. D., & Lawley, T. D. (2021). Massive expansion of human gut bacteriophage diversity. *Cell*, *184*(4), 1098–1109.e9. https://doi.org/10.1016/j.cell.2021.01.029

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics (Oxford, England)*, *34*(17), i884–i890. https://doi.org/10.1093/bioinformatics/bty560

Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., & de Hoon, M. J. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics (Oxford, England)*, *25*(11), 1422–1423. https://doi.org/10.1093/bioinformatics/btp163

Czepiel, J., Dróżdż, M., Pituch, H., Kuijper, E. J., Perucki, W., Mielimonka, A., Goldman, S., Wultańska, D., Garlicki, A., & Biesiada, G. (2019). Clostridium difficile infection: review. *European journal of clinical microbiology & infectious diseases : official publication of the European Society of Clinical Microbiology*, *38*(7), 1211–1221. https://doi.org/10.1007/s10096-019-03539-6

Dabrowska, K., Opolski, A., Wietrzyk, J., Switala-Jelen, K., Godlewska, J., Boratynski, J., Syper, D., Weber-Dabrowska, B., & Gorski, A. (2004). Anticancer activity of bacteriophage T4 and its mutant HAP1 in mouse experimental tumour models. *Anticancer research*, *24*(6), 3991–3995.

Dutilh, B. E., Cassman, N., McNair, K., Sanchez, S. E., Silva, G. G., Boling, L., Barr, J. J., Speth, D. R., Seguritan, V., Aziz, R. K., Felts, B., Dinsdale, E. A., Mokili, J. L., & Edwards, R. A. (2014). A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nature communications*, *5*, 4498. https://doi.org/10.1038/ncomms5498

Dion, M. B., Oechslin, F., & Moineau, S. (2020). Phage diversity, genomics and phylogeny. *Nature reviews. Microbiology*, *18*(3), 125–138. https://doi.org/10.1038/s41579-019-0311-5

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics (Oxford, England)*, *32*(19), 3047–3048. https://doi.org/10.1093/bioinformatics/btw354

Feiss, M., & Rao, V. B. (2012). The bacteriophage DNA packaging machine. *Advances in experimental medicine and biology*, *726*, 489–509. https://doi.org/10.1007/978-1-4614-0980-9_22

Frati, F., Salvatori, C., Incorvaia, C., Bellucci, A., Di Cara, G., Marcucci, F., & Esposito, S. (2018). The Role of the Microbiome in Asthma: The Gut-Lung Axis. *International journal of molecular sciences*, *20*(1), 123. https://doi.org/10.3390/ijms20010123

Garmaeva, S., Gulyaeva, A., Sinha, T., Shkoporov, A. N., Clooney, A. G., Stockdale, S. R., Spreckels, J. E., Sutton, T. D. S., Draper, L. A., Dutilh, B. E., Wijmenga, C., Kurilshikov, A., Fu, J., Hill, C., & Zhernakova, A. (2021). Stability of the human gut virome and effect of gluten-free diet. *Cell reports*, *35*(7), 109132. https://doi.org/10.1016/j.celrep.2021.109132

Ghurye, J., & Pop, M. (2019). Modern technologies and algorithms for scaffolding assembled genomes. *PLoS computational biology*, *15*(6), e1006994. https://doi.org/10.1371/journal.pcbi.1006994

Hatfull, G. F., & Hendrix, R. W. (2011). Bacteriophages and their genomes. *Current opinion in virology*, *1*(4), 298–303. https://doi.org/10.1016/j.coviro.2011.06.009

Hyatt, D., Chen, G. L., Locascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics*, *11*, 119. https://doi.org/10.1186/1471-2105-11-119

Islam, R., Raju, R. S., Tasnim, N., Shihab, I. H., Bhuiyan, M. A., Araf, Y., & Islam, T. (2021). Choice of assemblers has a critical impact on de novo assembly of SARS-CoV-2 genome and characterizing variants. *Briefings in bioinformatics*, *22*(5), bbab102. https://doi.org/10.1093/bib/bbab102

Lai, S., Pan, S., Sun, C., Coelho, L. P., Chen, W. H., & Zhao, X. M. (2022). metaMIC: reference-free misassembly identification and correction of de novo metagenomic assemblies. *Genome biology*, *23*(1), 242. https://doi.org/10.1186/s13059-022-02810-y

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Ma, Y., You, X., Mai, G., Tokuyasu, T., & Liu, C. (2018). A human gut phage catalog correlates the gut phageome with type 2 diabetes. *Microbiome*, *6*(1), 24. https://doi.org/10.1186/s40168-018-0410-y

Mikheenko, A., Saveliev, V., & Gurevich, A. (2016). MetaQUAST: evaluation of metagenome assemblies. *Bioinformatics (Oxford, England)*, *32*(7), 1088–1090. https://doi.org/10.1093/bioinformatics/btv697

Nurk, S., Meleshko, D., Korobeynikov, A., & Pevzner, P. A. (2017). metaSPAdes: a new versatile metagenomic assembler. *Genome research*, *27*(5), 824–834. https://doi.org/10.1101/gr.213959.116

Prjibelski, A. D., Vasilinetc, I., Bankevich, A., Gurevich, A., Krivosheeva, T., Nurk, S., Pham, S., Korobeynikov, A., Lapidus, A., & Pevzner, P. A. (2014). ExSPAnder: a universal repeat resolver for DNA fragment assembly. *Bioinformatics (Oxford, England)*, *30*(12), i293–i301. https://doi.org/10.1093/bioinformatics/btu266

Ren, J., Ahlgren, N. A., Lu, Y. Y., Fuhrman, J. A., & Sun, F. (2017). VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome*, *5*(1), 69. https://doi.org/10.1186/s40168-017-0283-5

Rinninella, E., Raoul, P., Cintoni, M., Franceschi, F., Miggiano, G. A. D., Gasbarrini, A., & Mele, M. C. (2019). What is the Healthy Gut Microbiota Composition? A Changing Ecosystem across Age, Environment, Diet, and Diseases. *Microorganisms*, *7*(1), 14. https://doi.org/10.3390/microorganisms7010014

Shkoporov, A. N., & Hill, C. (2019). Bacteriophages of the Human Gut: The "Known Unknown" of the Microbiome. *Cell host & microbe*, *25*(2), 195–209. https://doi.org/10.1016/j.chom.2019.01.017

Skorynina, A. V., Koposova, O. N., Kazantseva, O. A., Piligrimova, E. G., Ryabova, N. A., & Shadrin, A. M. (2022). Isolation and Characterization of Two Novel Siphoviruses Novomoskovsk and Bolokhovo, Encoding Polysaccharide Depolymerases Active against *Bacillus pumilus*. *International journal of molecular sciences*, *23*(21), 12988. https://doi.org/10.3390/ijms232112988

Tetz, G., Brown, S. M., Hao, Y., & Tetz, V. (2018). Parkinson's disease and bacteriophages as its overlooked contributors. *Scientific reports*, *8*(1), 10812. https://doi.org/10.1038/s41598-018-29173-4

Walker, P. J., Siddell, S. G., Lefkowitz, E. J., Mushegian, A. R., Adriaenssens, E. M., Alfenas-Zerbini, P., Dempsey, D. M., Dutilh, B. E., García, M. L., Curtis Hendrickson, R., Junglen, S., Krupovic, M., Kuhn, J. H., Lambert, A. J., Łobocka, M., Oksanen, H. M., Orton, R. J., Robertson, D. L., Rubino, L., Sabanadzovic, S., … Zerbini, F. M. (2022). Recent changes to virus taxonomy ratified by the International Committee on Taxonomy of Viruses (2022). *Archives of virology*, *167*(11), 2429–2440. https://doi.org/10.1007/s00705-022-05516-5

Wang, Q., Li, F., Liang, B., Liang, Y., Chen, S., Mo, X., Ju, Y., Zhao, H., Jia, H., Spector, T. D., Xie, H., & Guo, R. (2018). A metagenome-wide association study of gut microbiota in asthma in UK adults. *BMC microbiology*, *18*(1), 114. https://doi.org/10.1186/s12866-018-1257-x

Wilson, N. G., Hernandez-Leyva, A., Schwartz, D. J., Bacharier, L. B., & Kau, A. L. (2023). The gut metagenome harbors metabolic and antibiotic resistance signatures of moderate-to-severe asthma. *bioRxiv : the preprint server for biology*, 2023.01.03.522677. https://doi.org/10.1101/2023.01.03.522677

Zuo, T., Wong, S. H., Lam, K., Lui, R., Cheung, K., Tang, W., Ching, J. Y. L., Chan, P. K. S., Chan, M. C. W., Wu, J. C. Y., Chan, F. K. L., Yu, J., Sung, J. J. Y., & Ng, S. C. (2018). Bacteriophage transfer during faecal microbiota transplantation in *Clostridium difficile* infection is associated with treatment outcome. *Gut*, *67*(4), 634–643. https://doi.org/10.1136/gutjnl-2017-313952