viruses

Article

Exploring the Complexity of the Human Respiratory Virome through an In Silico Analysis of Shotgun Metagenomic Data Retrieved from Public Repositories

Talya Conradie [1,2], Jose A. Caparros-Martin [1], Siobhon Egan [2,3], Anthony Kicic [1,4,5,6], Sulev Koks [7,8], Stephen M. Stick [4,5] and Patricia Agudelo-Romero [1,9,10,*]

[1] Wal-Yan Respiratory Research Centre, Telethon Kids Institute, Perth, WA 6009, Australia
[2] Medical, Molecular and Forensic Sciences, Murdoch University, Perth, WA 6150, Australia
[3] Centre for Computational and Systems Medicine, Health Future Institute, Murdoch University, Perth, WA 6150, Australia
[4] Department of Respiratory and Sleep Medicine, Perth Children's Hospital for Children, Perth, WA 6009, Australia
[5] Centre for Cell Therapy and Regenerative Medicine, School of Medicine and Pharmacology, Perth, WA 6009, Australia
[6] School of Population Health, Curtin University, Perth, WA 6102, Australia
[7] Perron Institute for Neurological and Translational Science, Perth, WA 6009, Australia
[8] Centre for Molecular Medicine and Innovative Therapeutics, Murdoch University, Perth, WA 6150, Australia
[9] Australian Research Council Centre of Excellence in Plant Energy Biology, School of Molecular Sciences, The University of Western Australia, Perth, WA 6009, Australia
[10] European Virus Bioinformatics Centre, Friedrich-Schiller-Universitat Jena, 07737 Jena, Germany
* Correspondence: patricia.agudeloromero@telethonkids.org.au

Abstract: Background: Respiratory viruses significantly impact global morbidity and mortality, causing more disease in humans than any other infectious agent. Beyond pathogens, various viruses and bacteria colonize the respiratory tract without causing disease, potentially influencing respiratory diseases' pathogenesis. Nevertheless, our understanding of respiratory microbiota is limited by technical constraints, predominantly focusing on bacteria and neglecting crucial populations like viruses. Despite recent efforts to improve our understanding of viral diversity in the human body, our knowledge of viral diversity associated with the human respiratory tract remains limited. Methods: Following a comprehensive search in bibliographic and sequencing data repositories using keyword terms, we retrieved shotgun metagenomic data from public repositories (n = 85). After manual curation, sequencing data files from 43 studies were analyzed using EVEREST (pipEline for Viral assEmbly and chaRactEriSaTion). Complete and high-quality contigs were further assessed for genomic and taxonomic characterization. Results: Viral contigs were obtained from 194 out of the 868 FASTQ files processed through EVEREST. Of the 1842 contigs that were quality assessed, 8% (n = 146) were classified as complete/high-quality genomes. Most of the identified viral contigs were taxonomically classified as bacteriophages, with taxonomic resolution ranging from the superkingdom level down to the species level. Captured contigs were spread across 25 putative families and varied between RNA and DNA viruses, including previously uncharacterized viral genomes. Of note, airway samples also contained virus(es) characteristic of the human gastrointestinal tract, which have not been previously described as part of the lung virome. Additionally, by performing a meta-analysis of the integrated datasets, ecological trends within viral populations linked to human disease states and their biogeographical distribution along the respiratory tract were observed. Conclusion: By leveraging publicly available repositories of shotgun metagenomic data, the present study provides new insights into viral genomes associated with specimens from the human respiratory tract across different disease spectra. Further studies are required to validate our findings and evaluate the potential impact of these viral communities on respiratory tract physiology.

check for updates

Citation: Conradie, T.; Caparros-Martin, J.A.; Egan, S.; Kicic, A.; Koks, S.; Stick, S.M.; Agudelo-Romero, P. Exploring the Complexity of the Human Respiratory Virome through an In Silico Analysis of Shotgun Metagenomic Data Retrieved from Public Repositories. *Viruses* **2024**, *16*, 953. https://doi.org/10.3390/v16060953

Academic Editor: Hernan Garcia-Ruiz

Received: 21 March 2024
Revised: 31 May 2024
Accepted: 5 June 2024
Published: 13 June 2024

Keywords: respiratory viruses; shotgun metagenomics; lung virome; genome assembly; viromics; microbiome; airway; omics

1. Introduction

The human body hosts a diverse array of colonizing microorganisms, including bacteria, fungi, archaea, protozoa, and viruses, collectively forming ecological communities referred to as the human microbiota [1]. The composition of the human microbiota varies between different organ systems, with inter-individual differences arising from various factors, such as environmental exposure and genetic influences [1]. The observed variability in the microbiota's functional and taxonomic composition has also been linked to some diseases causally associated with various pathogenic processes in humans [2]. However, due to the simplicity of profiling bacteria using universal marker genes, much of the research on the human microbiota has predominantly focused on its bacterial component, with other components, such as viruses, remaining largely understudied [3].

The term "virome" refers to a subpopulation of the microbiota, which includes different types of viruses that can be classified depending on their host. This includes viruses that infect eukaryotic cells, and bacteriophages, which are capable of infecting bacteria [4–6]. Both eukaryotic viruses and bacteriophages can impact human health in both harmful and beneficial ways, including triggering immune responses or influencing residential bacterial communities within the same compartments [4,6]. In the last few decades, emerging infectious diseases caused by viruses, such as severe acute respiratory syndrome coronavirus (SARS-CoV-1), H1N1 influenza virus (Swine flu), Middle Eastern respiratory syndrome coronavirus (MERS-CoV), and severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), have become increasingly recognized as major threats to public health. Therefore, cataloguing existing viromes and understanding host–virus interactions are of essential value to protecting both individual and population level health [4,6–8].

While the lungs were initially considered a sterile organ, emerging evidence indicates the presence of a distinct microbiota [9]. The lower airways host a transient microbial population that likely originates in the oronasal cavity and reaches the lungs via micro-aspiration [10]. As a result, it is probable that respiratory epithelia are continually exposed to various clinically relevant viruses [9,11]. Viruses from multiple families, including bacteriophages, have been identified within the respiratory tract [4]. Certain viral families have shown associations with specific clinical groups; for instance, bacteriophages are frequently found in samples from individuals with cystic fibrosis, and *Anelloviridae* are prevalent in specimens from lung transplant recipients [5,7]. Eukaryotic viruses impact human infections and their progression, while bacteriophages can shape resident bacterial populations and alter functions of bacterial genomes by facilitating DNA transfer between cells, contributing to the spread of antimicrobial resistance [4,12,13].

Current lung virome studies are often subjected to unstandardized methodologies, leading to difficulties in viral identification. Up to 95% of sequences in different studies remain uncharacterized due to methodological variations [4,5,7,14–16]. This uncharacterized portion is referred to as "viral dark matter" and can be attributed to the incapability of databases that were generated using limited shotgun metagenomic data and often fail to assign sequences to known taxonomy. With metagenomic sequencing technologies rapidly improving, large datasets are being generated. However, computational strategies and current reference databases have not expanded at the same pace, limiting the ability to characterize viral sequences [17,18]. It is widely accepted that our current understanding of viral classification and taxonomy greatly underestimates the true diversity of viruses. The main public repository for viral sequences is the National Centre for Biotechnology (NCBI) database, but it tends to be biased as it does not encompass the full diversity of the viral community. Additionally, available reference sequences are biased towards those that can be cultivated and are clinically relevant, narrowing the scope of identifiable viruses.

This leads to purely novel and unidentified viruses remaining uncharacterized, and those slightly different from reference genomes potentially being mischaracterized [17,18]. Before exploring new samples and datasets, revisiting existing databases and thoroughly characterizing the available sample sets can expand diversity and reinforce current databases. This strategy has been employed in various human body sites, such as the skin and the gut, to address the issue of "viral dark matter" [19–21]. With lung virome research lagging in terms of available material, accessing and reanalyzing historical data are critical for a better understanding of true lung virome diversity. The aim of this study was to collate publicly available sequence data from human respiratory samples and reanalyze them with the objective of providing a more comprehensive understanding of viral diversity associated with the respiratory tract.

2. Materials and Methods

2.1. Bibliographic Search and Data Collection

Two strategies were used to collate lung virome studies for which sequencing data was publicly available (Figure 1). Strategy 1 involved a literature search on PubMed, utilizing specific keywords related to "lungs", "sputum" or "lavage fluid", "next generation sequencing" or "metagenomics", "bioinformatics", and "lung virome" (Table S1), and identified 18 studies with associated sequencing data obtained from the indicated repositories (Figure 2). Strategy 2 involved exploring sequence data repositories from NCBI Sequence Read Archive (SRA), NCBI Gene Expression Omnibus (GEO), and EMBL-EBI European Nucleotide Archive (ENA), using specific keywords related to "lung virome", "bacteriophages", "metagenomics", and "shotgun sequencing" to capture bioprojects that may contain unique information not yet linked to a published study. This second strategy yielded 67 unique studies (Figure 2).

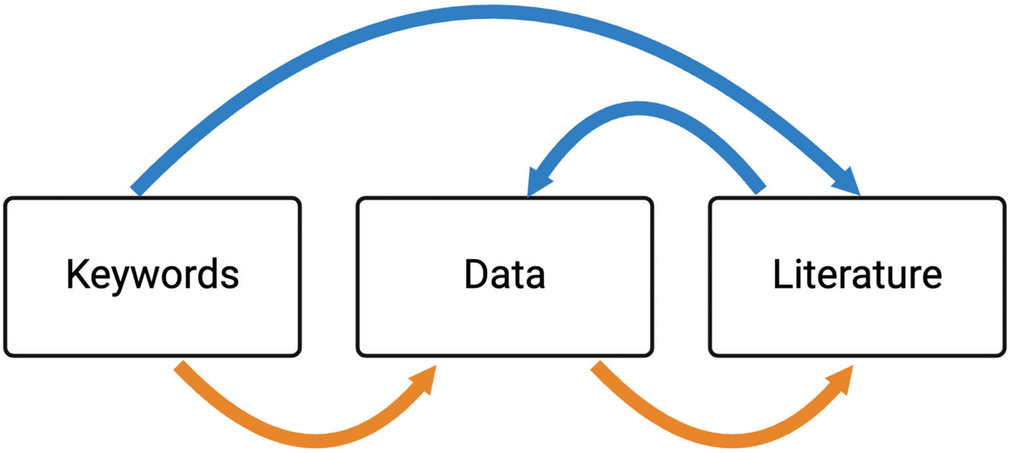

Figure 1. The bibliographic search and data collection strategy used. A flow diagram depicting the two search strategies used for data collection, with the blue arrows representing search strategy 1, and orange arrows representing search strategy 2.

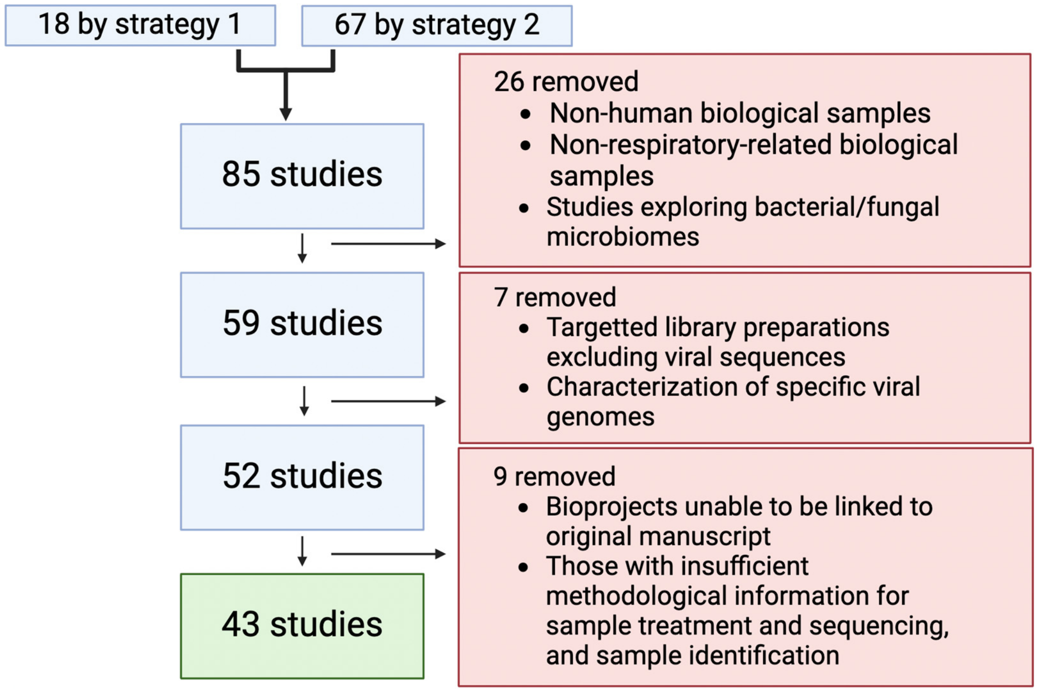
18 by strategy 1
67 by strategy 2
85 studies
26 removed
• Non-human biological samples
• Non-respiratory-related biological samples
• Studies exploring bacterial/fungal microbiomes
59 studies
7 removed
• Targetted library preparations excluding viral sequences
• Characterization of specific viral genomes
52 studies
9 removed
• Bioprojects unable to be linked to original manuscript
• Those with insufficient methodological information for sample treatment and sequencing, and sample identification
43 studies

Figure 2. The study-filtering strategy used. The flow diagram graphically represents the search strategy employed to identify relevant sequence data from public repositories. The number of studies excluded at each filtering step and the reasons for excluding these bioprojects are also indicated. A final 43 studies were identified that matched the selection criteria and were used for analysis in the present study.

2.2. Curation of Study Information

A total of 85 bioprojects were identified and linked to their original published studies, which were used to extract methodological information and metadata. Key study details, including title, authors, citation, year, and methodological specifics, were entered into our in-house database. Subsequently, a methodological screening of the database was conducted (Table S2).

Lung or respiratory biofluid samples were categorized as saliva, nasopharyngeal swabs and aspirates, sputum, bronchoalveolar lavage (BAL) fluid, and lung tissue. Studies were excluded under the following conditions: those using samples of non-human origin, biological material unrelated to the lung/respiratory tract, studies focused solely on the bacterial microbiome or fungal mycobiome using targeted amplicon or enrichment library preparation methods that would exclude viral sequences, studies focused on characterizing specific viral genomes, or those lacking identification in the original manuscript. Only those studies linked to their original papers or those providing sufficient methodological information in bioproject-associated metadata were retained. Ultimately, 43 studies met the screening criteria and were further analyzed to establish a curated database of the airway virome (Figure 2).

Each of the bioprojects associated with these 43 studies was classified based on study design, sequencing technology, study purpose (diseases or healthy clinical phenotypes), subject age, sample type, and sample size (Table S2). Additionally, the following were considered: sequencing platforms (e.g., short-read or long-read), types of reads (single- and/or paired-end), sequencing strategy (DNA, RNA, or both), and the nucleic acid extraction kits and library preparation techniques used. These methodological features were sourced from the materials and methods sections and the supplementary data within each study. In cases where discrepancies were identified between the information in the peer-reviewed published article and the sequence repository, the data available in the sequence repository were prioritized.

2.3. Retrieving Raw FASTQ Files

FASTQ files were downloaded from specific repositories using the software fastq-dl (v1.0.6, accessed on 1 April 2022, https://github.com/rpetit3/fastq-dl). A total of 43 bioprojects containing 868 independent sequencing files were downloaded.

Issues were identified with six of the studies from the finalized database, which included data not being available for download, data not correctly linked to the respective bioproject, or the absence of FASTQ files for download despite their stated availability [22–27]. To address this, formal requests were made via email to both the SRA, and the first and corresponding authors of the respective bioprojects. As a result, two bioprojects were successfully incorporated into the database. In one case, the SRA released the data for a bioproject [22], and in another, a linkage error was corrected [23]. Additionally, although two bioprojects were not accessible, the authors of those projects provided direct access to their data [24,25].

2.4. Bioinformatic Analysis

Data were processed using the bioinformatic pipeline EVEREST (https://github.com/agudeloromero/EVEREST, v0.01, accessed on 1 April 2022). EVEREST is an end-to-end pipeline designed for virus discovery, structured into five main phases that use FASTQ files as input. Briefly, during the pre-processing phase, files undergo quality control through trimming [26,27], followed by a filtering phase that includes host removal, replicated sequences elimination, and digital normalization [28–30]. Next, a de novo assembly is constructed using SPAdes [31] and similar contigs are clustered [32]. In the refinement phase, viral contigs are captured with VirSorter2 [33] and their quality is assessed using CheckV [34]. Finally, during the viral classification phase, two databases are used, nucleotide (NCBI) and amino acid (Uniprot), to taxonomically classify the viral contigs [32,35]. Each process is executed by select and specific software tools organized within the pipeline itself, as illustrated in Figure S1.

The data output produced by EVEREST was used for taxonomic classification. Each individual FASTQ file within a bioproject generated its own output files. Therefore, data was isolated and compiled separately for each sequencing file and then aggregated for each bioproject. Subsequently, these bioproject-level datasets were consolidated into a single metadata file.

The compiled data included both qualitative and quantitative information. Qualitative data included taxonomic and Baltimore classification, classification as a provirus, and quality levels as per CheckV [34]. Quantitative information included metrics such as GC content, contig length, and RPKM (Read Per Kilobase of reference sequence per Million total sequencing reads).

2.5. Statistical Analysis

Statistical analysis was performed in R v4.0.2. The normality of data distribution was evaluated using a Shapiro–Wilk test, and a Wilcoxon rank sum test was performed to compare groups when normality assumptions were not met. When conducting multiple comparisons, type I errors were controlled for by adjusting *p*-values using the Bonferroni correction method. Principal component analysis models were generated using a mixOmics R package [36]. Ecological alpha diversity estimates and PERMANOVA analysis were performed using the functions of R package Vegan (https://github.com/vegandevs/vegan, v2.6-4, accessed on 30 May 2024). The cut-off for statistical significance was set at *p*-value < 0.05.

3. Results

3.1. Study and Data Demographics

Following rigorous screening of the current literature and public repositories, a total of 85 lung virome metagenome studies were compiled, of which 43 were selected after manual curation (Tables S2 and S3). Out of the 43 bioprojects, 30 contained paired-end sequencing data and were processed through our bioinformatic pipeline, EVEREST, with

16 providing results (Tables 1, S2 and S3). The remaining 13 projects utilized a single-end sequencing strategy and were not processed because, at the time of the study, EVEREST was not equipped to handle this sequencing format. Among the 30 reanalyzed bioprojects, there were a total of 868 FASTQ sequencing files. EVEREST successfully detected viral contigs in 194 (22%) of these FASTQ files, which were associated with 16 different bioprojects (refer to Data Availability Statement, Tables 1, S2 and S3).

Table 1. General description of the bioprojects from which EVEREST successfully recovered viral contigs. A more comprehensive data table is provided in Table S3.

Bioproject	Clinical Group	Library Preparation	Sample Location	Total FASTQ Files	Successful FASTQ Files
PRJEB32062 [37]	CF	DNA	Sputum	25	18
PRJEB7454 [38]	TB	DNA	Mummified lung tissue	9	6
PRJNA316588 [39]	CF, COPD, Healthy smokers	DNA	Sputum	18	11
PRJNA369654 [40]	LTR, HIV	RNA and DNA	BAL	22	14
PRJNA392272 [41]	Sarcoidosis	RNA and DNA	BAL	98	25
PRJNA419524 [14]	LTR	RNA and DNA	BAL	63	40
PRJNA493096 [42]	Respiratory failure	RNA	BAL	4	2
PRJNA494633 [43]	ARI	RNA and DNA	NPS, NPA, Sputum	39	32
PRJNA573045 [44]	URTI	RNA	NPS	4	4
PRJNA601736 [45]	COVID-19	RNA	BAL	2	2
PRJNA623895 [46]	COVID-19	RNA	NPS	1	1
PRJNA629087 [47]	LTR	RNA	BAL	21	1
PRJNA671740 [48]	Lung adenocarcinoma	RNA and DNA	LT	5	1
PRJNA779483 [49]	ARI	RNA and DNA	Throat swabs	37	33
PRJNA189842 [50]	TB	DNA	Mummified lung tissue	1	1
PRJNA639353 [51]	Healthy	RNA	NTS	91	3

ARI, acute respiratory infection; CF, cystic fibrosis; COPD, chronic obstructive pulmonary disease; HIV, human immunodeficiency virus; LTR, lung transplant recipient; TB, tuberculosis; URTI, upper respiratory tract infection; BAL, bronchoalveolar lavage; LT, lung tissue; NPS, nasopharyngeal swabs; NPA, nasopharyngeal aspirates; NTS, nasal-throat swabs.

The 194 FASTQ files with successful viral contigs represented 523 subjects, including 131 adults, with 84 of them presenting a respiratory pathology, and 392 children, with 384 of them diagnosed with acute respiratory infection (Tables 1, S2 and S3, Figure S2A). In some studies, individual FASTQ files represented pooled samples from different individuals, which explains the discrepancy between the sequencing files and number of subjects represented in those files. The 523 individuals represent 13 different groups based on clinical phenotypes provided in the manuscript, including different respiratory pathologies, such as acute respiratory infection, cystic fibrosis, sarcoidosis, and lung transplant recipients (Tables 1, S2 and S3, Figure S2B). The studies represented a range of biological specimens from the upper and lower respiratory tract, with bronchoalveolar lavage (BAL) fluid, sputum, throat swabs, and nasopharyngeal aspirates proving to be the most successful sample types for viral contig detection (Tables 1, S2 and S3, Figure S2C).

3.2. Viral Contig Classification

EVEREST identified a total of 1842 contigs based on quality, of which 1414 were assigned taxonomy. Among these, 146 (8%) were classified as proviruses and excluded from further analysis, while the remaining 1696 (92%) were classified as viruses (Figure S3). EVEREST annotated 1193 (84%) contigs as bacteriophages, and the remaining 221 (16%) were identified as eukaryotic viral contigs (Figure 3A). Contigs were also grouped following the Baltimore classification system [52]. Most of the bacteriophage contigs (610, 51%) were classified as Group I, corresponding to double-stranded DNA bacteriophages. Six bacteriophage contigs (0.5%) were classified as Group II, representing single-stranded DNA bacteriophages. The remaining bacteriophage contigs could not be classified within the Baltimore system as they were not classified taxonomically to the family level. Eukaryotic viral contigs were distributed across four specific Baltimore groups, with the highest proportion (83, 37%) identified as Group I, followed by Group II (43, 20%), Group IV (20, 9%), and Group V (7, 3%) (Figure 3B).

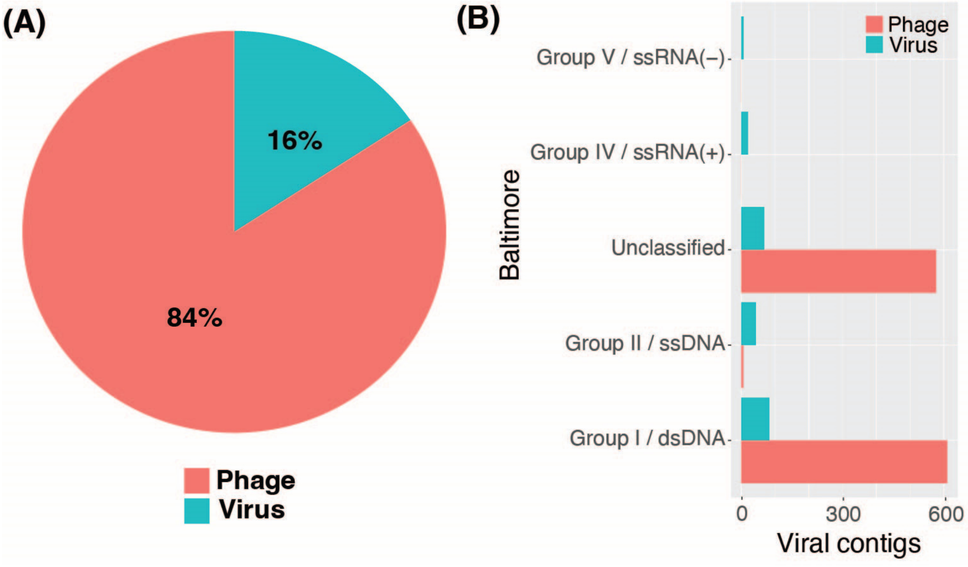

Figure 3. The distribution of eukaryotic viral contigs and bacteriophage contigs captured through EVEREST. (**A**) A pie chart depicting the 1414 captured contigs assigned to taxonomy, from which 1193 (84%) were identified as bacteriophages (red), and 221 (16%) were identified as eukaryotic viruses (blue). (**B**) Baltimore classification of viral contigs. The bar plot summarizes the number of eukaryote virus (blue) and bacteriophage (red) contigs within the indicated Baltimore classes, including those that could not be classified (unclassified).

Bacteriophage taxonomy was annotated to the order level (513, 43%), family level (366, 30%), subfamily rank (13, 1%), genus level (64, 5%), and species level (71, 6%). The remaining bacteriophage contigs were assigned to higher taxonomic ranks, including superkingdom (64, 5%) and clade (15, 1%), with 87 contigs (7%) unassigned to taxonomy (Figure 4A). Eukaryotic viral contigs were classified at the superkingdom (47, 21%), family (51, 23%), genus (36, 16%), or species (46, 21%) level. The remaining contigs were annotated at different taxonomic ranks or unassigned (15, 7%) (Figure 4A).

Viral contigs annotated at least down to the family level were distributed across 25 different families, including one unclassified family. Bacteriophages were distributed across 11 families, with the majority classified into the families *Siphoviridae* (359, 30%), *Myoviridae* (103, 9%), or *Podoviridae* (24, 2%). EVEREST also annotated 27 contigs as CrAssphages or CrAss-like phages (Table S4), a type of bacteriophage commonly observed in human fecal metagenomes [53]. These contigs were associated with six FASTQ files from three different bioprojects (PRJNA392272, PRJNA419524, and PRJNA494633) (Table S4) [14,41,43]. Eukaryotic viral contigs were distributed across 19 families, including *Phycodnaviridae* (18, 8%), *Siphoviridae* (17, 8%), *Herpesviridae* (16, 7%), *Microviridae* (13, 6%), *Papillomaviridae* (13, 6%), *Picornaviridae* (12, 5%), *Anelloviridae* (10, 4%), and *Papillomaviridae* (10, 4%) (Figure 4B).

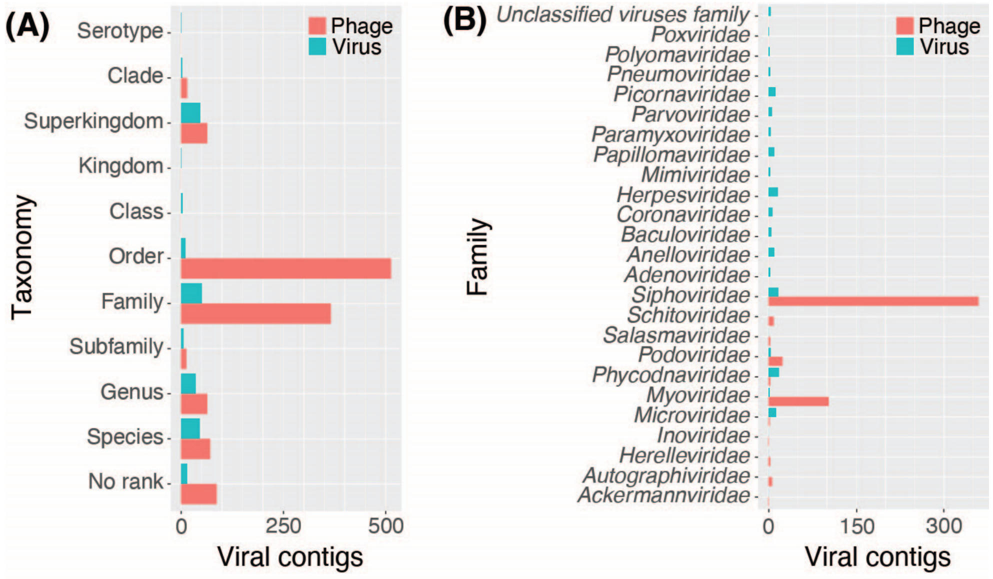

Figure 4. The distribution of taxonomic classification following the lowest common ancestor method included in EVEREST. (**A**) A bar plot showing the number of eukaryotic (blue) and bacteriophage (red) viral contigs classified at the indicated taxonomy ranks. (**B**). Taxonomic classification at the family level. The bar plots represent the 25 families identified among the bacteriophage (red) and eukaryotic (blue) viral contigs.

3.3. *Assessment of Viral and Proviral Genomes*

EVEREST implements CheckV [34] to evaluate the quality of the assembled contigs and ranked them based on their completeness in comparison to assessed reference genomes. The 1842 identified contigs were categorized into five quality tiers [34], with the majority being classified as low quality (0–50% completeness; 1111, 60%) or undetermined quality (for which a completeness estimate was available; 448, 24%). The remaining contigs were identified as medium-quality (50–90% completeness; 137, 7%), high-quality (>90% completeness 83, 4%), and complete (63, 3%) genomes (Figure 5A).

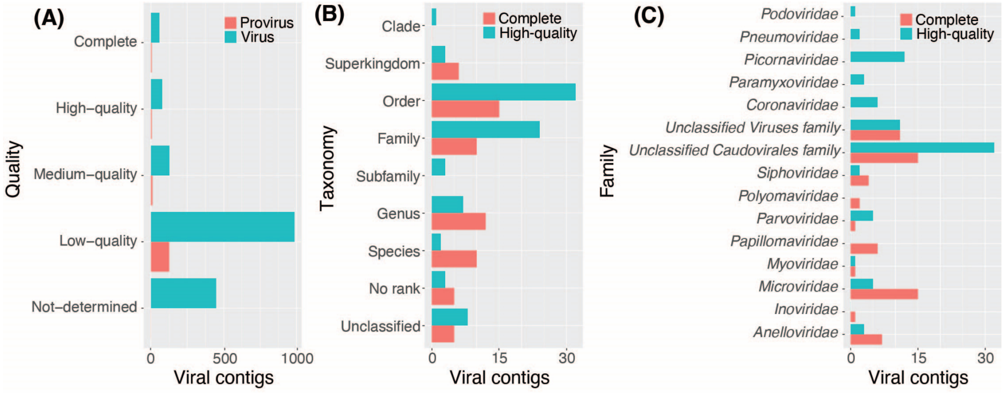

Figure 5. Genome quality assessment of the viral contigs and taxonomic characterization of the complete and high-quality genomes. (**A**) Bar plots represent the distribution of the 1842 contigs classified based on their quality as determined by CheckV [34]. Blue bars represent viral genomes, while red bars represent proviruses. (**B**) Bar plots representing the number of complete (red) and high-quality (blue) genomes that were assigned at each taxonomic rank. (**C**) Bar plots representing the number of complete (red) or high-quality (blue) viral genomes that were assigned to the indicated viral families.

Among the 146 contigs classified as complete or high-quality genomes, most were annotated up to order (47, 32%) and family (34, 23%) levels, with 21 (14%) annotated as either unclassified or no rank (Figure 4B). Complete and high-quality contigs were assigned to 15 different viral families, with most of these contigs annotated as unclassified families within the Caudovirales order (47, 32%), followed by *Microviridae* (20, 14%), while 22 (15%) contigs were unclassified (Figure 5C).

To confirm that there was no bias in contig assembly associated with the different bioprojects, GC content and length were evaluated for the assembled genome for the 1414 eukaryotic and bacteriophage viral contigs assigned to taxonomy down to the family level (Figure 6). Contigs were distributed in a bimodal distribution regarding their GC content, with values ranging from 0.2 to 0.7. The contig length distribution was right-skewed, with values ranging from 5000 to 130,000 bp (Figure 6). Contigs annotated to specific families were clustered around similar GC content values, while contig length was neither associated with specific viral families nor linked to altered GC values, suggesting no bias in contig generation across different bioprojects.

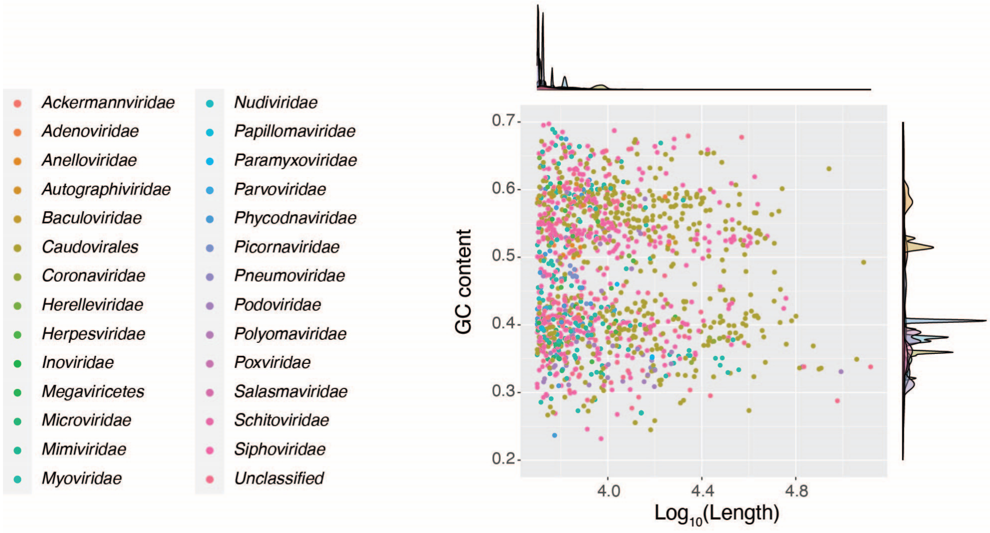

Figure 6. An overview of the viral contigs recovered from the bioprojects analyzed in this study grouped by taxonomic viral families. The scatter plot shows the correlation between GC content and viral contig size. Density plots of viral families on the top and right sides represent contig size and GC content, respectively.

For GC content, compared to families with a GC content under 0.4, some families were found to cluster in uniform sections, showing a uniform GC content of these viral families even between samples (Figure 6). In terms of genome length, some families showed strong and partial clustering, indicating that these families showed consistency in the length of their genomes between various studies and cohorts (Figure 6).

3.4. Metanalysis of Viral Communities in the Respiratory Tract

We next explored the viral community profiles obtained through EVEREST in relation to covariates, such as disease state, respiratory niche, or the type of biospecimen used to sample the airways. We excluded from these analyses the profiles obtained from the two reports studying 18th-century mummified lung tissue from subjects infected with *Mycobacterium tuberculosis* as the representativeness of viral particles and nucleic acids in these samples may be compromised due to their age of over 200 years [38,50]. Principal component analysis (PCA) was used to explore viral profiles in the samples included in this study for which viral contigs were obtained (Figure 7). The first two components of the PCA model explained 30% of the variance associated with our dataset, with an unclassified family within Caudovirales and *Siphoviridae* being the major drivers of variation along the first two components (Figure 7A).

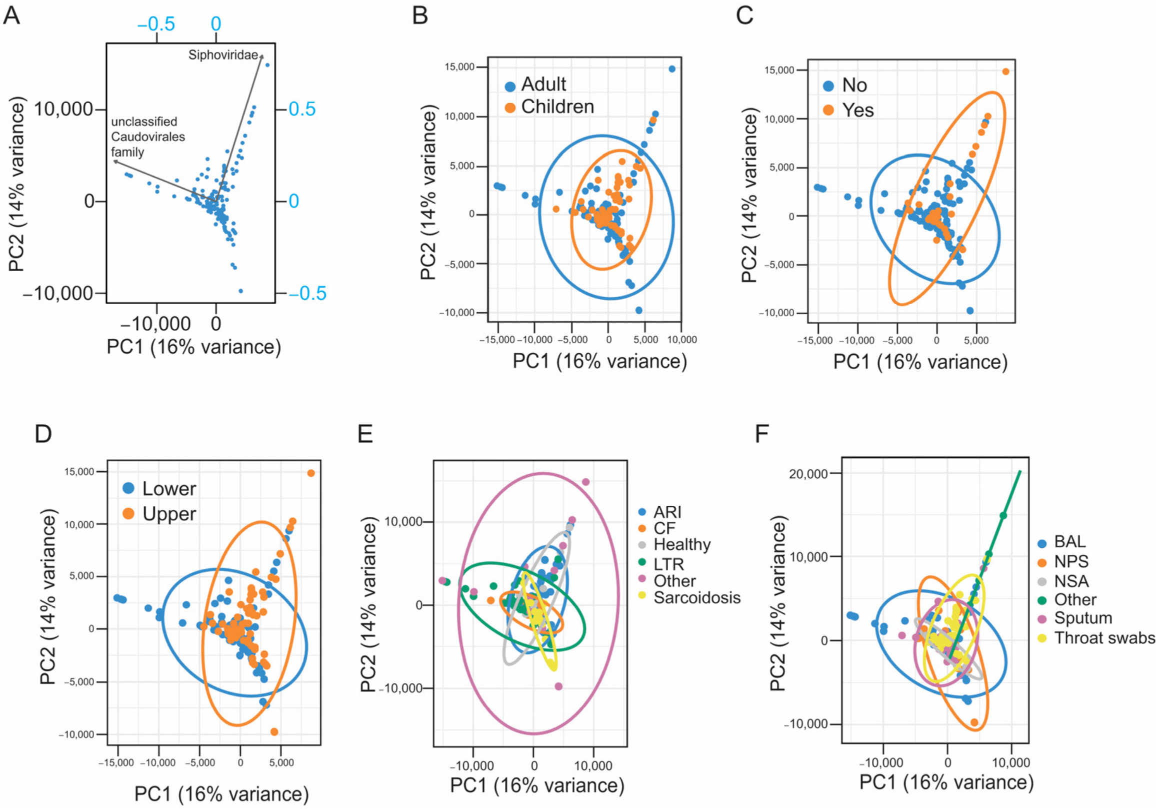

Figure 7. (**A**–**F**) A projection of the samples included in this study in the two first components of the PCA model. (**A**) A biplot showing the PCA score plot and the loading plot. The left and bottom axes represent the PCA scores of the samples, and the top and right axes indicate how strong each feature (vector) influences the principal components. Vectors represent features with at least a 0.6 correlation coefficient with any of the two first components of the PCA model, and their project values on each principal component indicates how much weight they have on those specific components of the PCA model. The unclassified family within Caudovirales is negatively correlated with component 1 and represents the major source of variation along this component. Conversely, *Siphoviridae* exhibit a positive correlation with component 1. Distribution across component 2 is mostly driven by *Siphoviridae* (positive correlation) and to a lesser extent by the unclassified family within Caudovirales. For each PCA graph, dots represent individual samples. Samples are labelled based on whether they were obtained from adults (blue) or children (orange) (**B**), whether they were obtained from healthy individuals (orange) or individuals with disease (blue) (**C**), whether they were obtained from the lower (blue) or upper (orange) respiratory tract (**D**), the specific respiratory pathology (**E**), or the type of specimen used to sample the airways (**F**). ARI: acute respiratory infection; CF: cystic fibrosis; LTR: lung transplant recipient; BAL: bronchoalveolar lavage; NPS: nasopharyngeal swabs; NSA: nasopharyngeal aspirates.

In our dataset, samples obtained from adults were associated with both an unclassified family within Caudovirales and *Siphoviridae* (Figure 7A,B). Interestingly, samples obtained from children clustered around the center of the PCA, indicating that they were not associated with the major source of variation represented by components 1 and 2 (Figure 7A,B). Likewise, samples from healthy individuals were associated with *Siphoviridae*, while those with a respiratory pathology showed overall positive correlation with an unclassified family within Caudovirales and negative correlation with *Siphoviridae* (Figure 7A,C). On the other hand, samples from the lower airways were mostly associated with an unclassified family within Caudovirales, while samples from the upper airways were mainly associated with the family *Siphoviridae* (Figure 7A,D). Variations in the virome of lung transplant recipients were associated with members of an unclassified family within Caudovirales (Figure 7A,E).

Likewise, the abundance of members from an unclassified family within Caudovirales was associated with the virome of BAL samples (Figure 7A,F).

To understand the impact of the defined variables, including age, disease type, and biospecimen type, on the composition of viral communities represented in the PCA model, a permutational multivariate analysis of variance (PERMANOVA) analysis was performed. Our findings revealed that the origin of the samples, denoted by the bioproject from which the samples were obtained, had a substantial effect on the overall viral community structure represented in the PCA model (PERMANOVA, *pseudo* F = 3.90, R^2 = 0.23, *p*-value = 0.0001). Likewise, based on the sum of squares (10.9% explained by the type of sample, and 22.8% explained by the disease group), the results of the PERMANOVA analysis suggest that, within our dataset, both biospecimen type (PERMANOVA, *pseudo* F = 3.51, R^2 = 0.11, *p* = 0.0002) and disease groups (PERMANOVA, *pseudo* F = 4.50, R^2 = 0.23, *p*-value = 0.0001) had a strong effect over the observed structure of the viral communities associated with the respiratory tract. Interestingly, when consolidating the different disease groups into a single "disease" category, the impact of the binary variable, disease or healthy, on the overall structure of the respiratory virome in our dataset was less pronounced compared to the analysis that considered the individual disease states (PERMANOVA, *pseudo* F = 2.93, R^2 = 0.02, *p*-value = 0.003), perhaps suggesting disease-specific viral community profiles. Similarly, other predefined groups, such as patient age categories (adults or children) and respiratory tract niche (upper or lower) associated with each of the respiratory samples incorporated into the PCA model, had a small effect on the overall structure of the resulting viral communities (PERMANOVA for age, *pseudo* F = 4.01, R^2 = 0.02, *p*-value = 0.0001; PERMANOVA for respiratory niche, *pseudo* F = 5.71, R^2 = 0.03, *p* = 0.0001).

The alpha diversity estimate, as measured by the Shannon index, was assessed and compared across different sample-associated descriptors, including biospecimen origin, age group category, and clinical phenotype, using pairwise comparisons (Figure 8). For this analysis, we excluded viral profiles from bioprojects PRJNA419524 and PRJEB32062 [14,37] as they represent longitudinal samples and were, therefore, analyzed separately. Biospecimen origin, whether from the upper or lower respiratory tract, showed no effect on viral diversity (Wilcoxon rank sum test (WRST), *p*-value = 0.09) (Figure 8A). Similarly, age group category was not associated with viral diversity (WRST, *p*-value = 0.06) (Figure 8B). Conversely, samples collected from healthy individuals demonstrated higher diversity than those subjects diagnosed with a respiratory pathology (WRST, *p*-value = 0.0007) (Figure 8C). To gain further insights into this association, viral diversity in the context of specific disease conditions was explored. Compared to healthy controls, acute respiratory infections (WRST with Bonferroni correction, *p*-value = 0.004) and lung transplant recipients (WRST with Bonferroni correction, *p*-value = 0.002) were associated with lower viral diversity in the respiratory tract (Figure 8D). On the contrary, patients with sarcoidosis demonstrated comparable diversity values to that of healthy controls (Figure 8D). Lastly, the impact of the type of sampling on viral diversity was examined (Figure 8E). Compared to bronchoalveolar lavage (BAL) fluid, viral diversity in nasopharyngeal aspirates (WRST with Bonferroni correction, *p*-value = 0.04), sputum (WRST with Bonferroni correction, *p*-value = 0.02), and throat swab samples (WRST with Bonferroni correction, *p*-value = 0.002) were higher (Figure 8E). In contrast, viral diversity in nasopharyngeal swabs was comparable to that of BAL fluid (Figure 8E).

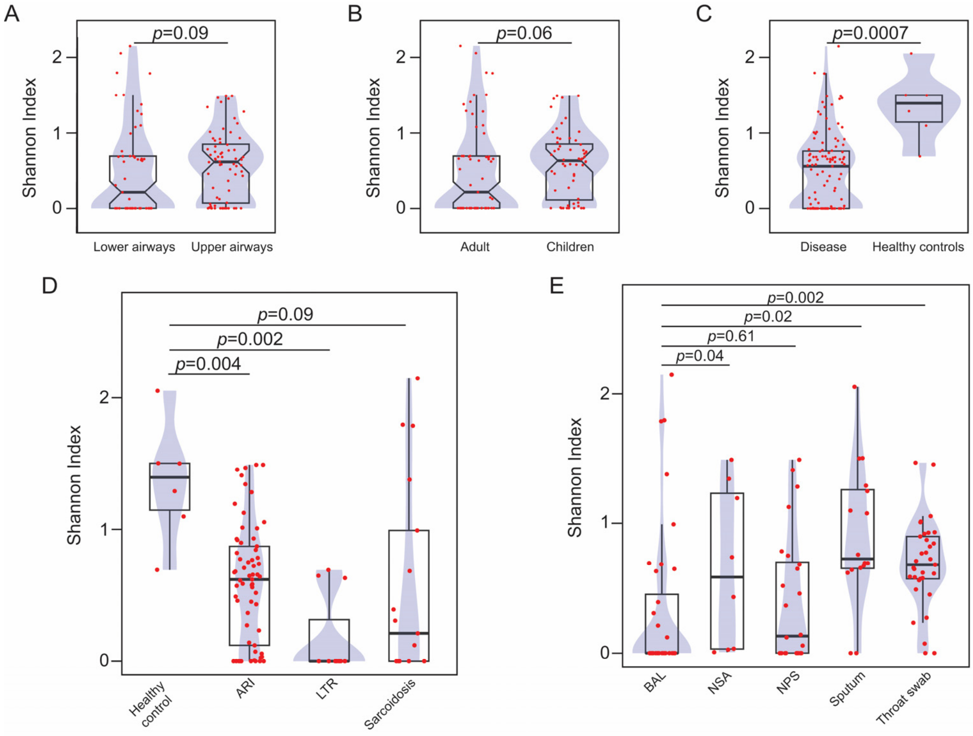

Figure 8. Shannon index diversity of viral families identified in lung samples. (**A**–**E**) Box plots displaying the comparison in alpha viral diversity between samples from the upper and lower respiratory tract (**A**), between samples collected from adults and children (**B**), between samples obtained from healthy controls or individuals with disease (**C**), between samples obtained from patients with different respiratory pathologies and healthy controls (**D**), and between bronchoalveolar lavage fluid and other type of specimens from the respiratory tract (**E**). Red dots represent individual datapoints. Only groups containing more than three datapoints were included in these analyses. The density plots (shaded blue area) represent the probability distribution for each data group. Differences between the indicated groups were evaluated using a Wilcoxon rank sum test and the resulting *p*-values are indicated in each panel. In A and B, notches in the boxplot represent the 95% confidence interval for the median. In D and E, *p*-values were corrected using the Bonferroni method. ARI: acute respiratory infection; LTR: lung transplant recipient; BAL: bronchoalveolar lavage; NSA: nasopharyngeal aspirates; NPS: nasopharyngeal swabs.

Bioproject PRJEB32062 contains sequencing files representing longitudinal sputum samples from patients with cystic fibrosis, which were obtained during both exacerbation episodes and periods of clinical stability [37]. After running EVEREST, we obtained longitudinal viral profiles from three patients. As shown in Figure S4, viral diversity in these samples was highly variable over time, with exacerbation episodes associated with both high and low viral diversity values even in samples obtained from the same patient (Figure S4). Bioproject PRJNA419524 contains sequencing files from BAL fluid obtained from donor lungs and recipients, the latter being monitored longitudinally [14]. EVEREST reported viral profiles from 40 FASTQ files within this dataset representing 13 LTR subjects (Tables S3 and S4). Longitudinal data were available for 11 LTR subjects, of which 7 LTR subjects also had viral profiles from the donor lungs (Tables S3 and S4). As shown in Figure S5, viral alpha diversity was largely equivalent between BAL collected from donor lungs and the first BAL collected post-transplant from the recipient (mean difference 0.39, 95% confidence interval [−0.54, 1.32]) (Figure S5A). Likewise, we did not observe differences in BAL-associated viral diversity between the first two consecutive timepoints collected post-transplant (mean difference 0.15, 95% confidence interval [−0.42, 0.72]) (Figure S5B).

4. Discussion

In this study, we reanalyzed sequence data files from biological specimens obtained from the human respiratory tract to better understand viral composition and diversity associated with the human airways. Although similar studies have been performed in samples from the human gut and skin niches [19–21], our study represents a new insight into understanding viral diversity associated with the respiratory system in humans. Our bioinformatic pipeline recovered 1842 viral contigs, with 146 representing either complete or near-complete viral genomes. Taxonomy was assigned to 1414 contigs (77%) representing 25 viral families. This classification included both known and unidentified viral families within the human respiratory tract. The analysis of identified contigs unveiled a wide spectrum of viral families, including some complete and high-quality genomes. In general, EVEREST reproduced the viral profiles reported in the original studies, except for one study in which EVEREST did not capture contigs classified within the *Anelloviridae* family. However, while the original study accessed the NCBI database in 2013 [40], EVEREST uses a more updated version [35], which might explain these discrepancies. Interestingly, most of the recovered contigs were classified as bacteriophages (Table S4). This observation is in line with the observed high diversity of bacteriophages in other human niches such as the gut, in which phages might play an important role in the modulation of the intestinal ecosystem [54]. A more plausible explanation for the high observed diversity of bacteriophages in the reanalyzed studies/datasets is the presence of a specific bacterial host. Accordingly, we observed *Mycobacterium* phages in FASTQ files from both bioprojects containing data from individuals positive for *Mycobacterium* infection (PRJEB7454 and PRJNA189842) [38,50]. Likewise, in bioprojects containing sequencing data from people with cystic fibrosis such as PRJEB32062, phages were detected from pathogens typically found in the airways of patients with cystic fibrosis, such as *Pseudomonas*, *Staphylococcus*, or *Ralstonia* (Table S4) [37]. Notably, a viral family not previously identified within the lung virome was discovered. Thus, we recovered multiple genomes representing CrAssphages, a group of bacteriophages typically associated with the human gut [19,53,54]. This was an unexpected finding, given that previous lung virome studies have not identified CrAssphages in their datasets [16,26,40,55]. However, the presence of CrAssphage contigs in six different FASTQ files from three different studies [14,41,43] involving samples from both adults and children across various clinical groups strongly suggests that they do not represent contaminating sequences or methodological artifacts specific to individual studies. Although the biological significance of CrAssphages in the airways is not within the scope of our study, these viruses may enter the airways along with gastrointestinal bacteria through reflux-micro-aspiration processes, which are commonly associated with different respiratory pathologies [56]. Overall, our observations suggest that, as it has been recently observed in the gut [19,54], bacteriophage populations constitute a significant component of human respiratory microbiota.

Ecological descriptors were used to assess viral community profiles in relation to covariates, such as disease states, different respiratory niches, or the type of biospecimen used to survey the airways. Most viral community profile variation was explained by the different bioprojects from which the sequencing data were obtained. This finding is not surprising as methodological differences in obtaining viral nucleic acids, such as the introduction of viral enrichment steps, or other aspects, such as sequencing library preparations or sequencing methods, are expected to substantially influence the portion of the virome that will be captured in each study [57]. Similarly, the type of specimen used for sampling the airways and the disease group from which the biological specimen was obtained significantly impacted the resulting viral community profiles. The influence of the type of specimen on viral community composition may be attributed to niche-specific bacterial communities along the respiratory tract [58] or the impact of a higher proportion of host DNA on sequencing sensitivity for virome profiling (e.g., when using lung tissue samples) [59]. Thus, nasopharyngeal aspirates, sputum, and throat swab samples demonstrated higher viral diversity compared to BAL fluid, which showed similar diversity to nasopharyngeal swab samples. On the other hand, variation in viral community profiles associated with different clinical groups may be indicative of disease-specific airway bacterial compositions. For example, viruses belonging to the family *Siphoviridae* were correlated with the virome of healthy subjects, while an unclassified family within the Caudovirales order was associated with lung transplant recipients. Interestingly, while individuals with disease demonstrated reduced viral diversity in the airways compared to healthy controls, this trend was not consistently applicable across the different pathologies evaluated in our study. In contrast to samples obtained from acute respiratory infections or lung transplant recipients, which displayed lower viral diversity in comparison to samples from healthy controls, alpha viral diversity estimates in specimens from patients with sarcoidosis were equivalent to those of healthy individuals. Furthermore, compared to the viral profiles observed in adult samples, the viral communities in samples obtained from children were not associated with specific viral families, even though 98% of these samples were obtained from children with acute respiratory infections. This observation suggests that, as with the bacterial component of the microbiota, there could be either an age-related ecological succession in the viral communities of the human respiratory tract [58,60] or the airway virome in children is more resilient to perturbation.

Our study has several limitations. Firstly, the small number of studies and datasets available at the time of analysis, coupled with the highly heterogenous population related to both the clinical presentation and respiratory niche sampled, represent challenges. Thus, it is important to exercise caution when interpreting the findings of our study as the limited number of available studies and differences in airway sampling methodologies across the various cohorts could have potentially confounded our observations. Furthermore, most of the analyzed studies targeted DNA, with only one study targeting both RNA and DNA populations. Consequently, our study is biased towards DNA viruses, likely underestimating the true diversity of the virome associated with the human respiratory tract.

5. Conclusions

In conclusion, our exploratory analysis provides new insights into the viral families present in the human respiratory tract and offers preliminary observations regarding covariates that may influence viral community composition. By leveraging publicly available repositories of sequencing data, the present study provides a more nuanced picture of the viral population associated with the human airways and illustrates how the implementation of a recently developed computational pipeline can be used to characterize viral sequences from shotgun metagenomic sequencing data repositories.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/v16060953/s1: Figure S1: Diagram depicting the EVEREST pipelines; Figure S2: Demographics of the studies analyzed in this study; Figure S3: Pie chart showing the distribution of contigs classified as viruses and proviruses; Figure S4: Scatter plot showing the variation in viral diversity over time in sputum from three patients with cystic fibrosis (bioproject PRJEB32062); Figure S5: Boxplots showing alpha diversity estimates (Shannon index) in paired BAL samples; Table S1: Search terms utilized in strategy 1 and 2 for bibliographic data search and data collection; Table S2: Summary of the 43 bioprojects and their associated studies; Table S3; Related to Table 1. Detailed metadata linked to the FASTQ files from which EVEREST classified viral contigs; Table S4: Taxonomic paths and RPKM data output from EVEREST for every FASTQ file from which EVEREST detected viral contigs.

Author Contributions: Conceptualization: P.A.-R. and J.A.C.-M.; methodology: P.A.-R. and T.C.; formal analysis: P.A.-R., T.C. and J.A.C.-M.; investigation: T.C.; resources: P.A.-R.; supervision: P.A.-R., J.A.C.-M., S.E., A.K., S.K. and S.M.S.; writing–review and editing: P.A.-R., T.C., J.A.C.-M., S.E., A.K., S.K. and S.M.S.; visualization: P.A.-R., T.C. and J.A.C.-M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Australian Government Research Training Program Scholarship, the National Health and Medical Research Council of Australia (NHMRC, grant number

NHMRC 115648), the Australian NHMRC 2020 Synergy grant (grant number APP118364). A.K. is a Rowell Family Fellow; S.M.S. is supported by an NHMRC Investigator Award (2021/GR000216). P.A.-R. received funding from a Google Cloud Education Program grant and a Telethon Kids Institute Theme Collaboration Award 2023.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The following bioprojects were successfully inputted through EVEREST: PRJEB32062 (https://www.ncbi.nlm.nih.gov/bioproject/PRJEB32062/) (accessed on 1 April 2022), PRJEB7454 (https://www.ncbi.nlm.nih.gov/bioproject/PRJEB7454/) (accessed on 1 April 2022), PRJNA316588 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA316588/) (accessed on 1 April 2022), PRJNA369654 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA369654/) (accessed on 1 April 2022), PRJNA392272 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA392272/) (accessed on 1 April 2022), PRJNA419524 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA419524/) (accessed on 1 April 2022), PRJNA493096 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA493096/) (accessed on 1 April 2022), PRJNA494633 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA494633/) (accessed on 1 April 2022), PRJNA573045 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA573045/) (accessed on 1 April 2022), PRJNA601736 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA601736/) (accessed on 1 April 2022), PRJNA623895 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA623895/) (accessed on 1 April 2022), PRJNA629087 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA629087/) (accessed on 1 April 2022), PRJNA671740 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA671740/) (accessed on 1 April 2022), PRJNA779483 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA779483/) (accessed on 1 April 2022), PRJNA189842 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA189842/) (accessed on 1 April 2022), PRJNA639353 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA639353/) (accessed on 1 April 2022). All bioprojects from which the experimental data were retrieved are presented in Supplementary Table S2, with the bioproject accession number, reference, and study information and citation [14,22,23,37–51,61–85].

Acknowledgments: We would like to acknowledge the Pawsey Supercomputing Research Centre and the Australian BioCommons through the Australian BioCommons Leadership Share (ABLeS) project for supporting with computing resources.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Gilbert, J.A.; Blaser, M.J.; Caporaso, J.G.; Jansson, J.K.; Lynch, S.V.; Knight, R. Current understanding of the human microbiome. *Nat. Med.* **2018**, *24*, 392–400. [CrossRef] [PubMed]
2. Hou, K.; Wu, Z.X.; Chen, X.Y.; Wang, J.Q.; Zhang, D.; Xiao, C.; Zhu, D.; Koya, J.B.; Wei, L.; Li, J.; et al. Microbiota in health and diseases. *Signal Transduct. Target. Ther.* **2022**, *7*, 135. [CrossRef] [PubMed]
3. Thomas, A.M.; Segata, N. Multiple levels of the unknown in microbiome research. *BMC Biol.* **2019**, *17*, 48. [CrossRef] [PubMed]
4. Liang, G.; Bushman, F.D. The human virome: Assembly, composition and host interactions. *Nat. Rev. Microbiol.* **2021**, *19*, 514–527. [CrossRef] [PubMed]
5. Mitchell, A.B.; Oliver, B.G.G.; Glanville, A.R. Translational Aspects of the Human Respiratory Virome. *Am. J. Respir. Crit. Care Med.* **2016**, *194*, 1458–1464. [CrossRef] [PubMed]
6. Tan, S.K.; Relman, D.A.; Pinsky, B.A. The Human Virome: Implications for Clinical Practice in Transplantation Medicine. *J. Clin. Microbiol.* **2017**, *55*, 2884–2893. [CrossRef] [PubMed]
7. Zou, S.; Caler, L.; Colombini-Hatch, S.; Glynn, S.; Srinivas, P. Research on the human virome: Where are we and what is next. *Microbiome* **2016**, *4*, 32. [CrossRef] [PubMed]
8. Jones, K.E.; Patel, N.G.; Levy, M.A.; Storeygard, A.; Balk, D.; Gittleman, J.L.; Daszak, P. Global trends in emerging infectious diseases. *Nature* **2008**, *451*, 990–993. [CrossRef] [PubMed]
9. Dickson, R.P.; Erb-Downward, J.R.; Martinez, F.J.; Huffnagle, G.B. The Microbiome and the Respiratory Tract. *Annu. Rev. Physiol.* **2016**, *78*, 481–504. [CrossRef]
10. Dickson, R.P.; Erb-Downward, J.R.; Freeman, C.M.; McCloskey, L.; Falkowski, N.R.; Huffnagle, G.B.; Curtis, J.L. Bacterial Topography of the Healthy Human Lower Respiratory Tract. *mBio* **2017**, *8*, e02287. [CrossRef]
11. Mitchell, A.B.; Glanville, A.R. Introduction to Techniques and Methodologies for Characterizing the Human Respiratory Virome. *Methods Mol. Biol.* **2018**, *1838*, 111–123. [PubMed]

12. Klumpp, J.; Fouts, D.E.; Sozhamannan, S. Next generation sequencing technologies and the changing landscape of phage genomics. *Bacteriophage* **2012**, *2*, 190–199. [CrossRef] [PubMed]
13. Pfeifer, E.; Bonnin, R.A.; Rocha, E.P.C. Phage-Plasmids Spread Antibiotic Resistance Genes through Infection and Lysogenic Conversion. *mBio* **2022**, *13*, e0185122. [CrossRef] [PubMed]
14. Abbas, A.A.; Young, J.C.; Clarke, E.L.; Diamond, J.M.; Imai, I.; Haas, A.R.; Cantu, E.; Lederer, D.J.; Meyer, K.; Milewski, R.K.; et al. Bidirectional transfer of anelloviridae lineages between graft and host during lung transplantation. *Am. J. Transplant.* **2019**, *19*, 1086–1097. [CrossRef] [PubMed]
15. Xu, L.; Zhu, Y.; Ren, L.; Xu, B.; Liu, C.; Xie, Z.; Shen, K. Characterization of the nasopharyngeal viral microbiome from children with community-acquired pneumonia but negative for Luminex xTAG respiratory viral panel assay detection. *J. Med. Virol.* **2017**, *89*, 2098–2107. [CrossRef] [PubMed]
16. Willner, D.; Furlan, M.; Haynes, M.; Schmieder, R.; Angly, F.E.; Silva, J.; Tammadoni, S.; Nosrat, B.; Conrad, D.; Rohwer, F. Metagenomic analysis of respiratory tract DNA viral communities in cystic fibrosis and non-cystic fibrosis individuals. *PLoS ONE* **2009**, *4*, e7370. [CrossRef] [PubMed]
17. Rose, R.; Constantinides, B.; Tapinos, A.; Robertson, D.L.; Prosperi, M. Challenges in the analysis of viral metagenomes. *Virus Evol.* **2016**, *2*, vew022. [CrossRef]
18. Kieft, K.; Anantharaman, K. Virus genomics: What is being overlooked? *Curr. Opin. Virol.* **2022**, *53*, 101200. [CrossRef] [PubMed]
19. Camarillo-Guerrero, L.F.; Almeida, A.; Rangel-Pineros, G.; Finn, R.D.; Lawley, T.D. Massive expansion of human gut bacteriophage diversity. *Cell* **2021**, *184*, 1098–1109.e9. [CrossRef]
20. Modha, S.; Robertson, D.L.; Hughes, J.; Orton, R.J. Quantifying and Cataloguing Unknown Sequences within Human Microbiomes. *mSystems* **2022**, *7*, e0146821. [CrossRef]
21. Elbehery, A.H.A.; Feichtmayer, J.; Singh, D.; Griebler, C.; Deng, L. The human virome protein cluster database (HVPC): A human viral metagenomic database for diversity and function annotation. *Front. Microbiol.* **2018**, *9*, 1110. [CrossRef] [PubMed]
22. Goolam Mahomed, T.; Peters, R.P.H.; Allam, M.; Ismail, A.; Mtshali, S.; Goolam Mahomed, A.; Ueckermann, V.; Kock, M.M.; Ehlers, M.M. Lung microbiome of stable and exacerbated COPD patients in Tshwane, South Africa. *Sci. Rep.* **2021**, *11*, 19758. [CrossRef] [PubMed]
23. Wang, Y.; Zhu, N.; Li, Y.; Lu, R.; Wang, H.; Liu, G.; Zou, X.; Xie, Z.; Tan, W. Metagenomic analysis of viral genetic diversity in respiratory samples from children with severe acute respiratory infection in China. *Clin. Microbiol. Infect.* **2016**, *22*, 458.e1–458.e9. [CrossRef] [PubMed]
24. Einarsson, G.G.; Vanaudenaerde, B.M.; Spence, C.D.; Lee, A.J.; Boon, M.; Verleden, G.M.; Elborn, J.S.; Dupont, L.J.; Van Raemdonck, D.; Gilpin, D.F.; et al. Microbial Community Composition in Explanted Cystic Fibrosis and Control Donor Lungs. *Front. Cell Infect. Microbiol.* **2022**, *11*, 764585. [CrossRef]
25. Dinsdale, E.A.; Edwards, R.A.; Hall, D.; Angly, F.; Breitbart, M.; Brulc, J.M.; Furlan, M.; Desnues, C.; Haynes, M.; Li, L.; et al. Functional metagenomic profiling of nine biomes. *Nature* **2008**, *452*, 629–632. [CrossRef]
26. Andrews, S.; Krueger, F.; Segonds-Pichon, A.; Biggins, F.; Fastqc, W.S. *A Quality Control Tool for High Throughput Sequence Data*; Babraham Bioinformatics, Babraham Institute: Cambridge, UK, 2015. Available online: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed on 23 September 2022).
27. Bolger, A.M.; Lohse, M.; Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **2014**, *30*, 2114–2120. [CrossRef]
28. Li, H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **2018**, *34*, 3094–3100. [CrossRef] [PubMed]
29. Li, H.; Handsaker, B.; Wysoker, A.; Fennel, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R.; 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and SAMtools. *Bioinformatics* **2009**, *25*, 2078–2079. [CrossRef] [PubMed]
30. Bushnel, I.B. BBMAP: Short-Read Aligner, and Other Bioinformatics Tools. 2016. Available online: http://sourceforge.net/projects/bbmap/ (accessed on 23 September 2022).
31. Prjibelski, A.; Antipov, D.; Meleshko, D.; Lapidus, A.; Korobeynikov, A. Using SPAdes De Novo assembler. *Curr. Protoc. Bioinforma* **2020**, *70*, e102. [CrossRef]
32. Mirdita, M.; Steinegger, M.; Breitwieser, F.; Soding, J.; Levy Karin, E. Fast and sensitive taxonomic assignment to metagenomic contigs. *Bioinformatics* **2021**, *37*, 3029–3031. [CrossRef]
33. Guo, J.; Bolduc, B.; Zayed, A.A.; Varsani, A.; Dominguez-Huerta, G.; Delmont, T.O.; Pratama, A.A.; Gazitua, M.C.; Vik, D.; Sullivan, M.B.; et al. VirSorter2: A multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome* **2021**, *9*, 37. [CrossRef] [PubMed]
34. Nayfach, S.; Camargo, A.P.; Schulz, F.; Eloe-Fadrosh, E.; Roux, S.; Kyrpides, N.C. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat. Biotechnol.* **2021**, *39*, 578–585. [CrossRef] [PubMed]
35. Agudelo-Romero, P.; Sharma, A.; Conradie, T.; Kicic, A.; Caparros-Martin, J.A.; Stick, S.M. Database for EVEREST (Pipeline for Viral Assembly and chaRactEriSaTion) (0.03) [Data set]. *Zenodo*. Available online: https://zenodo.org/records/8404860 (accessed on 23 September 2022).
36. Rohart, F.; Gautier, B.; Singh, A.; Lê Cao, K.A. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS Comput. Biol.* **2017**, *13*, e1005752. [CrossRef] [PubMed]

37. Dmitrijeva, M.; Kahlert, C.R.; Feigelman, R.; Kleiner, R.L.; Nolte, O.; Albrich, W.C.; Baty, F.; von Mering, C. Strain-Resolved Dynamics of the Lung Microbiome in Patients with Cystic Fibrosis. *mBio* **2021**, *12*, e02863. [CrossRef] [PubMed]
38. Kay, G.L.; Sergeant, M.J.; Zhou, Z.; Chan, J.Z.M.; Millard, A.; Quick, J.; Szikossy, I.; Pap, I.; Spigelman, M.; Loman, N.J.; et al. Eighteenth-century genomes show that mixed infections were common at time of peak tuberculosis in Europe. *Nat. Commun.* **2015**, *6*, 6717. [CrossRef] [PubMed]
39. Feigelman, R.; Kahlert, C.R.; Baty, F.; Rassouli, F.; Kleiner, R.L.; Kohler, P.; Brutsche, M.H.; von Mering, C. Sputum DNA sequencing in cystic fibrosis: Non-invasive access to the lung microbiome and to pathogen details. *Microbiome* **2017**, *5*, 20. [CrossRef] [PubMed]
40. Young, J.C.; Chehoud, C.; Bittinger, K.; Bailey, A.; Diamond, J.M.; Cantu, E.; Haas, A.R.; Abbas, A.; Frye, L.; Christie, J.D.; et al. Viral metagenomics reveal blooms of anelloviruses in the respiratory tract of lung transplant recipients. *Am. J. Transplant.* **2015**, *15*, 200–209. [CrossRef]
41. Clarke, E.L.; Lauder, A.P.; Hofstaedter, C.E.; Hwang, Y.; Fitzgerald, A.S.; Imai, I.; Biernat, W.; Rekawiecki, B.; Majewska, H.; Dubaniewicz, A.; et al. Microbial Lineages in Sarcoidosis. A Metagenomic Analysis Tailored for Low–Microbial Content Samples. *Am. J. Respir. Crit. Care Med.* **2018**, *197*, 225–234. [CrossRef] [PubMed]
42. Mayday, M.Y.; Khan, L.M.; Chow, E.D.; Zinter, M.S.; Derisi, J.L. Miniaturization and optimization of 384-well compatible RNA sequencing library preparation. *PLoS ONE* **2019**, *14*, e0206194. [CrossRef]
43. Bal, A.; Pichon, M.; Picard, C.; Casalegno, J.S.; Valette, M.; Schuffenecker, I.; Billard, L.; Vallet, S.; Vilchez, G.; Cheynet, V.; et al. Quality control implementation for universal characterization of DNA and RNA viruses in clinical respiratory samples using single metagenomic next-generation sequencing workflow. *BMC Infect. Dis.* **2018**, *18*, 537. [CrossRef]
44. Hoque, M.N.; Rahman, M.S.; Ahmed, R.; Hossain, M.S.; Islam, M.S.; Islam, T.; Hossain, M.A.; Siddiki, A.Z. Diversity and genomic determinants of the microbiomes associated with COVID-19 and non-COVID respiratory diseases. *Gene Rep.* **2021**, *23*, 101200. [CrossRef] [PubMed]
45. Chen, L.; Liu, W.; Zhang, Q.; Xu, K.; Ye, G.; Wu, W.; Sun, Z.; Liu, F.; Wu, K.; Zhong, B.; et al. RNA based mNGS approach identifies a novel human coronavirus from two individual pneumonia cases in 2019 Wuhan outbreak. *Emerg. Microbes Infect.* **2020**, *9*, 313–319. [CrossRef] [PubMed]
46. Manning, J.E.; Bohl, J.A.; Lay, S.; Chea, S.; Sovann, L.; Sengdoeurn, Y.; Heng, S.; Vuthy, C.; Kalantar, K.; Ahyong, V.; et al. Rapid metagenomic characterization of a case of imported COVID-19 in Cambodia. *bioRxiv* **2020**, *5*, 2020.03.02.968818.
47. Mitchell, A.B.; Li, C.X.; Oliver, B.G.G.; Holmes, E.C.; Glanville, A.R. High-resolution Metatranscriptomic Characterization of the Pulmonary RNA Virome After Lung Transplantation. *Transplantation* **2021**, *105*, 2546–2553. [CrossRef] [PubMed]
48. Cai, H.Z.; Zhang, H.; Yang, J.; Zeng, J.; Wang, H. Preliminary assessment of viral metagenome from cancer tissue and blood from patients with lung adenocarcinoma. *J. Med. Virol.* **2021**, *93*, 5126–5133. [CrossRef] [PubMed]
49. Mao, Q.; Sun, G.; Qian, Y.; Qian, Y.; Li, W.; Wang, X.; Shen, Q.; Yang, S.; Zhou, C.; Wang, H.; et al. Viral metagenomics of pharyngeal secretions from children with acute respiratory diseases with unknown etiology revealed diverse viruses. *Virus Res.* **2022**, *321*, 198912. [CrossRef] [PubMed]
50. Chan, J.Z.M.; Sergeant, M.J.; Lee, O.Y.C.; Minnikin, D.E.; Besra, G.S.; Pap, I.; Spigelman, M.; Donoghue, H.; Pallen, M.J. Metagenomic Analysis of Tuberculosis in a Mummy. *N. Engl. J. Med.* **2013**, *369*, 289–290. [CrossRef]
51. Thi Kha Tu, N.; Thi Thu Hong, N.; Thi Han Ny, N.; My Phuc, T.; Thi Thanh Tam, P.; Doorn, H.R.V.; Dang Trung Nghia, H.; Thao Huong, D.; An Han, D.; Thi Thu Ha, L.; et al. The Virome of Acute Respiratory Diseases in Individuals at Risk of Zoonotic Infections. *Viruses* **2020**, *12*, 960. [CrossRef] [PubMed]
52. Baltimore, D. Expression of animal virus genomes. *Bacteriol. Rev.* **1971**, *35*, 235. [CrossRef]
53. Dutilh, B.E.; Cassman, N.; McNair, K.; Sanchez, S.E.; Silva, G.G.Z.; Boling, L.; Barr, J.J.; Speth, D.R.; Seguritan, V.; Aziz, R.K.; et al. A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat. Commun.* **2014**, *5*, 4498. [CrossRef]
54. Cao, Z.; Sugimura, N.; Burgermeister, E.; Ebert, M.P.; Zuo, T.; Lan, P. The gut virome: A new microbiome component in health and disease. *EBioMedicine* **2022**, *81*, 104113. [CrossRef] [PubMed]
55. Gregory, A.C.; Sullivan, M.B.; Segal, L.N.; Keller, B.C. Smoking is associated with quantifiable differences in the human lung DNA virome and metabolome. *Respir. Res.* **2018**, *19*, 174. [CrossRef]
56. Lee, A.S.; Lee, J.S.; He, Z.; Ryu, J.H. Reflux-Aspiration in Chronic Lung Disease. *Ann. Am. Thorac. Soc.* **2020**, *17*, 155–164. [CrossRef] [PubMed]
57. Parras-Moltó, M.; Rodríguez-Galet, A.; Suárez-Rodríguez, P.; López-Bueno, A. Evaluation of bias induced by viral enrichment and random amplification protocols in metagenomic surveys of saliva DNA viruses. *Microbiome* **2018**, *6*, 119. [CrossRef] [PubMed]
58. Man, W.H.; De Steenhuijsen Piters, W.A.A.; Bogaert, D. The microbiota of the respiratory tract: Gatekeeper to respiratory health. *Nat. Rev. Microbiol.* **2017**, *15*, 259–270. [CrossRef]
59. Pereira-Marques, J.; Hout, A.; Ferreira, R.M.; Weber, M.; Pinto-Ribeiro, I.; Van Doorn, L.J.; Knetsch, C.W.; Figueiredo, C. Impact of host DNA and sequencing depth on the taxonomic resolution of whole metagenome sequencing for microbiome analysis. *Front. Microbiol.* **2019**, *10*, 1277. [CrossRef]
60. Kumpitsch, C.; Koskinen, K.; Schöpf, V.; Moissl-Eichinger, C. The microbiome of the upper respiratory tract in health and disease. *BMC Biol.* **2019**, *17*, 87. [CrossRef] [PubMed]

61. Takayama, I.; Nguyen, B.G.; Dao, C.X.; Pham, T.T.; Dang, T.Q.; Truong, P.T.; Do, T.V.; Pham, T.T.P.; Fujisaki, S.; Odagiri, T.; et al. Next-generation sequencing analysis of the within-host genetic diversity of influenza A(H1N1)pdm09 viruses in the upper and lower respiratory tracts of patients with severe influenza. *mSphere* **2021**, *6*, e01043. [CrossRef] [PubMed]
62. de Castilhos, J.; Zamir, E.; Hippchen, T.; Rohrbach, R.; Schmidt, S.; Hengler, S.; Schumacher, H.; Neubauer, M.; Kunz, S.; Müller-Esch, T.; et al. COVID-19 severity and complications associated with low diversity, dysbiosis and predictive metagenome features of the oropharyngeal microbiome. *Res. Sq.* **2021**. [CrossRef]
63. Saito, T.; Miyagawa, K.; Chen, S.Y.; Tamosiuniene, R.; Wang, L.; Sharpe, O.; Samayoa, E.; Harada, D.; Moonen, J.R.A.J.; Cao, A.; et al. Upregulation of human endogenous retrovirus-K is linked to immunity and inflammation in pulmonary arterial hypertension. *Circulation* **2017**, *136*, 1920–1935. [CrossRef]
64. Bacci, G.; Mengoni, A.; Fiscarelli, E.; Segata, N.; Taccetti, G.; Dolce, D.; Paganin, P.; Morelli, P.; Tuccio, V.; De Alessandri, A.; et al. A different microbiome gene repertoire in the airways of Cystic Fibrosis patients with severe lung disease. *Int. J. Mol. Sci.* **2017**, *18*, 1654. [CrossRef] [PubMed]
65. Haswell, L.E.; Baxter, A.; Banerjee, A.; Verrastro, I.; Mushonganono, J.; Adamson, J.; Thorne, D.; Gaça, M.; Minet, E. Reduced biological effect of e-cigarette aerosol compared to cigarette smoke evaluated in vitro using normalized nicotine dose and RNA-seq-based toxicogenomics. *Sci. Rep.* **2017**, *7*, 888. [CrossRef] [PubMed]
66. Abbas, A.A.; Diamond, J.M.; Chehoud, C.; Chang, B.; Kotzin, J.J.; Young, J.C.; Imai, I.; Haas, A.R.; Cantu, E.; Lederer, D.J.; et al. The perioperative lung transplant virome: Torque Teno viruses are elevated in donor lungs and show divergent dynamics in primary graft dysfunction. *Am. J. Transplant.* **2017**, *17*, 1313–1324. [CrossRef]
67. Jaffe, D.; Muenzer, J.; Storch, G.; Weinstock, G.; Sodergren, E.; Wylie, K.; Arens, M.; Buller, R. The human virome in children and its relationship to febrile illness. *Nat. Preced.* **2010**. [CrossRef]
68. Bacci, G.; Taccetti, G.; Dolce, D.; Armanini, F.; Segata, N.; Di Cesare, F.; Lucidi, V.; Fiscarelli, E.; Morelli, P.; Casciaro, R.; et al. Untargeted metagenomic investigation of the airway microbiome of Cystic Fibrosis patients with moderate-severe lung disease. *Microorganisms* **2020**, *8*, 1003. [CrossRef]
69. Altan, E.; Dib, J.C.; Gulloso, A.R.; Juandigua, D.E.; Deng, X.; Bruhn, R.; Hildebrand, K.; Freiden, P.; Yamamoto, J.; Schultz-Cherry, S.; et al. Effect of geographic isolation on the nasal virome of indigenous children. *J. Virol.* **2019**, *93*, e00681-19. [CrossRef] [PubMed]
70. Tsitsiklis, A.; Osborne, C.M.; Kamm, J.; Williamson, K.; Kalantar, K.; Dudas, G.; Caldera, S.; Lyden, A.; Tan, M.; Neff, N.; et al. Lower respiratory tract infections in children requiring mechanical ventilation: A multicentre prospective surveillance study incorporating airway metagenomics. *Lancet Microbe.* **2022**, *3*, e284–e293. [CrossRef]
71. Van Rijn, A.L.; Van Boheemen, S.; Sidorov, I.; Carbo, E.C.; Pappas, N.; Mei, H.; Feltkamp, M.; Aanerud, M.; Bakke, P.; Claas, E.C.J.; et al. The respiratory virome and exacerbations in patients with chronic obstructive pulmonary disease. *PLoS ONE* **2019**, *14*, e0223952. [CrossRef]
72. Babiker, A.; Bradley, H.L.; Stittleburg, V.D.; Ingersoll, J.M.; Key, A.; Kraft, C.S.; Waggoner, J.J.; Piantadosi, A. Metagenomic sequencing to detect respiratory viruses in persons under investigation for COVID-19. *J. Clin. Microbiol.* **2020**, *59*, e02142-20. [CrossRef] [PubMed]
73. Pratas, D.; Toppinen, M.; Pyoria, L.; Hedman, K.; Sajantila, A.; Perdomo, M.F. A hybrid pipeline for reconstruction and analysis of viral genomes at multi-organ level. *Gigascience* **2020**, *9*, giaa086. [CrossRef] [PubMed]
74. Rajagopala, S.V.; Bakhoum, N.G.; Pakala, S.B.; Shilts, M.H.; Rosas-Salazar, C.; Mai, A.; Boone, H.H.; McHenry, R.; Yooseph, S.; Halasa, N.; et al. Metatranscriptomics to characterize respiratory virome, microbiome, and host response directly from clinical samples. *Cell Rep. Methods* **2021**, *1*, 100091. [CrossRef] [PubMed]
75. Welch, N.L.; Zhu, M.; Hua, C.; Weller, J.; Mirhashemi, M.E.; Nguyen, T.G.; Mantena, S.; Bauer, M.R.; Shaw, B.M.; Ackerman, C.M.; et al. Multiplexed CRISPR-based microfluidic platform for clinical testing of respiratory viruses and identification of SARS-CoV-2 variants. *Nat. Med.* **2022**, *28*, 1083–1094. [CrossRef] [PubMed]
76. Guo, Y.; Li, H.; Chen, H.; Li, Z.; Ding, W.; Wang, J.; Yin, Y.; Jin, L.; Sun, S.; Jing, C.; et al. Metagenomic next-generation sequencing to identify pathogens and cancer in lung biopsy tissue. *EBioMedicine* **2021**, *73*, 103639. [CrossRef] [PubMed]
77. Segura-Wang, M.; Görzer, I.; Jaksch, P.; Puchhammer-Stöckl, E. Temporal dynamics of the lung and plasma viromes in lung transplant recipients. *PLoS ONE* **2018**, *13*, e0200428. [CrossRef] [PubMed]
78. Garcia-Nuñez, M.; Gallego, M.; Monton, C.; Capilla, S.; Millares, L.; Pomares, X.; Espasa, M.; Ferrari, R.; Moya, A.; Monsó, E.; et al. The respiratory virome in chronic obstructive pulmonary disease. *Future Virol.* **2018**, *13*, 457–466. [CrossRef]
79. Xiao, Y.L.; Kash, J.C.; Beres, S.B.; Sheng, Z.M.; Musser, J.M.; Taubenberger, J.K. High-throughput RNA sequencing of a formalin-fixed, paraffin-embedded autopsy lung tissue sample from the 1918 influenza pandemic. *J. Pathol.* **2013**, *229*, 535–545. [CrossRef] [PubMed]
80. Hilton, S.K.; Castro-Nallar, E.; Pérez-Losada, M.; Toma, I.; McCaffrey, T.A.; Hoffman, E.P.; Siegel, M.O.; Simon, G.L.; Johnson, W.E.; Crandall, K.A. Metataxonomic and metagenomic approaches vs. culture-based techniques for clinical pathology. *Front. Microbiol.* **2016**, *7*, 484. [CrossRef] [PubMed]
81. Lee, S.W.; Kuan, C.S.; Wu, L.S.H.; Weng, J.T.Y. Metagenome and metatranscriptome profiling of moderate and severe COPD sputum in Taiwanese Han males. *PLoS ONE* **2016**, *11*, e0159066. [CrossRef] [PubMed]

82. Güemes, A.G.C.; Lim, Y.W.; Quinn, R.A.; Conrad, D.J.; Benler, S.; Maughan, H.; Edwards, R.; Brettin, T.; Cantú, V.A.; Cuevas, D.; et al. Cystic Fibrosis rapid response: Translating multi-omics data into clinically relevant information. *mBio* **2019**, *10*, e00431-19. [CrossRef]
83. Sabin, S.; Herbig, A.; Vågene, Å.J.; Ahlström, T.; Bozovic, G.; Arcini, C.; Kühnert, D.; Bos, K.I. A seventeenth-century Mycobacterium tuberculosis genome supports a Neolithic emergence of the Mycobacterium tuberculosis complex. *Genome Biol.* **2020**, *21*, 201. [CrossRef]
84. Lim, Y.W.; Schmieder, R.; Haynes, M.; Willner, D.; Furlan, M.; Youle, M.; Abbott, K.; Edwards, R.; Evangelista, J.; Conrad, D.; et al. Metagenomics and metatranscriptomics: Windows on CF-associated viral and microbial communities. *J. Cyst. Fibros.* **2013**, *12*, 154–164. [CrossRef] [PubMed]
85. Mokili, J.L.; Dutilh, B.E.; Lim, Y.W.; Schneider, B.S.; Taylor, T.; Haynes, M.R.; Metzgar, D.; Myers, C.A.; Blair, P.J.; Nosrat, B.; et al. Identification of a novel human papillomavirus by metagenomic analysis of samples from patients with febrile respiratory illness. *PLoS ONE* **2013**, *8*, e58404. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.