Total Infectomes Characterization of Respiratory Infections in pre-COVID-19 Wuhan, China

Mang Shi[1,2+], Su Zhao[3+], Bin Yu[4+], Wei-Chen Wu[1+], Yi Hu[3+], Jun-Hua Tian[4], Wen Yin[3], Fang Ni[3], Hong-Ling Hu[3], Shuang Geng[3], Li Tan[3], Ying Peng[4], Zhi-Gang Song[1], Wen Wang[1,5], Yan-Mei Chen[1], Edward C. Holmes[1,2], Yong-Zhen Zhang[1]*

[1]Shanghai Public Health Clinical Center, State Key Laboratory of Genetic Engineering, School of Life Sciences and Human Phenome Institute, Fudan University, Shanghai, China.

[2]Marie Bashir Institute for Infectious Diseases and Biosecurity, School of Life and Environmental Sciences and School of Medical Sciences, The University of Sydney, Sydney, Australia

[3]Department of Pulmonary and Critical Care Medicine, The Central Hospital of Wuhan, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China.

[4]Wuhan Center for Disease Control and Prevention, Wuhan, China.

[5]Department of Zoonosis, National Institute for Communicable Disease Control and Prevention, China Center for Disease Control and Prevention, Beijing, China.

+ Contributed equally to this article.

* Correspondence: zhangyongzhen@fudan.edu.cn

Abstract

At the end of 2019 Wuhan witnessed an outbreak of "atypical pneumonia" that later developed into a global pandemic. Metagenomic sequencing rapidly revealed the causative agent of this outbreak to be a novel coronavirus - SARS-CoV-2. Herein, to provide a snapshot of the pathogens in pneumonia-associated respiratory samples from Wuhan prior to the emergence of SARS-CoV-2, we collected bronchoalveolar lavage fluid samples from 408 patients presenting with pneumonia and acute respiratory infections at the Central Hospital of Wuhan between 2016 and 2017. Unbiased total RNA sequencing was performed to reveal their "total infectome", including viruses, bacteria and fungi. Consequently, we identified 37 pathogen species, comprising 15 RNA viruses, 3 DNA viruses, 16 bacteria and 3 fungi, often at high abundance and including multiple co-infections (12.8%). However, SARS-CoV-2 was not present. These data depict a stable core infectome comprising common respiratory pathogens such as rhinoviruses and influenza viruses, an atypical respiratory virus (EV-D68), and a single case of a sporadic zoonotic pathogen – *Chlamydia psittaci*. Samples from patients experiencing respiratory disease on average had higher pathogen abundance than healthy controls. Phylogenetic analyses of individual pathogens revealed multiple origins and global transmission histories, highlighting the connectedness of the Wuhan population. This study provides a comprehensive overview of the pathogens associated with acute respiratory infections and pneumonia, which were more diverse and complex than obtained using targeted PCR or qPCR approaches. These data also suggest that SARS-CoV-2 or closely related viruses were absent from Wuhan in 2016-2017.

Key words: Acute respiratory infection, pneumonia, total infectome, SARS-CoV-2, Wuhan

Introduction

The emergence of COVID-19 at the end of 2019 has had a profound impact on the world. The causative agent, a betacoronavirus (*Coronaviridae*) termed SARS-CoV-2, has high transmissibility, causes mild to severe respiratory symptoms in humans[1-3]. The first documented case of SARS-CoV-2 was reported in Wuhan, Hubei province, China[3]. Despite intensive research into SARS-CoV-2 and COVID-19, important questions remain unclear regarding the emergence of this virus, including for when and where it appeared and how long it was circulating in human populations prior to its initial detection in December 2019.

Meta-transcriptomics has several advantages over traditional diagnostic approaches based on serology or PCR[4]: it targets all types of micro-organisms simultaneously, identifies potential pathogens without *a priori* knowledge of what micro-organisms might be present, reveals the information (RNA) expressed by the pathogen during infection that is central to agent identification and studies of disease association. This method has been proven highly successful in revealing the entire virome and microbiome in a diverse range of species[5-8], including the initial identification of SARS-CoV-2 from patients with severe pneumonia[1].

Acute respiratory infections and pneumonia are a significant public health concern on a global scale. However, far less is known about the total "infectomes" associated with respiratory infections and pneumonia. Herein, we report total infectome surveillance of 408 patients presenting with pneumonia and acute respiratory infections at Wuhan Central Hospital prior to the SARS-CoV-2 epidemic. The purpose of this study was to use an un-biased meta-transcriptomics tool to characterize the total infectome within these patients. Nevertheless, since the sampling period occurred before the outbreak of COVID-19, this represents the first

opportunity to characterize the entire range of pathogens simultaneously within a cohort and determine the microbial composition of the population in which SARS-CoV-2 was initially reported.

Results

Patient context

We considered 408 patients clinically diagnosed with pneumonia or acute respiratory infection at the Central hospital of Wuhan in Wuhan, China. The sampling lasted for 20 months and covered the period between May 2016 to December 2017, two years before the onset of the COVID-19 pandemic (Fig. 1A). The male-to-female ratio among the patients sampled was 1.4 (Fig. 1B), with age ranging from 16 to 90 years (medium, 62). Pre-existing medical conditions present in these patients included hypertension (n=108), diabetes (n=46), bronchiectasia (n=31), chronic obstructive pulmonary disease (COPD, n=23), cancer (n=12), and heart disease (n=10). Based on evaluations made by clinicians at the hospital, 27 patients were described as severely ill, with 381 presenting with non-severe syndromes (Fig. 1E). The mortality for the entire cohort was 0.74% (n=3) and the average duration of hospitalization was 8 days (range 2-322, medium 9).

Total infectome

Meta-transcriptomic analysis of the BALF samples identified a wide range of RNA viruses, DNA viruses, bacteria and fungi. For the purposes of this study, we only characterized those likely associated with human disease (i.e., pathogens). This included (i) existing species that are known to be associated with human disease, and (ii) potentially novel pathogens that have not been previously characterized. For the latter, we only considered DNA and RNA viruses that are related to a virus genus or family that have previously been shown to infect mammals and are at

relatively high abundance level (i.e., >0.1% of total RNA, or 1000 RPM). Other than new pathogens, the abundance threshold for pathogen positives was set as 1 RPM. Furthermore, commensal bacteria population was not considered here.

Based on these criteria we did not identify any potential novel viral pathogens. All the microbes identified belonged to those previously characterized as human pathogens, comprising 15 RNA viruses, 3 DNA viruses, 16 bacteria and 3 fungal pathogens (Fig. 2). The case positive rate for all pathogens was 71.25% (n=249, Fig. 2A), many of which were only associated with RNA viruses (27.3%, n=151) or bacteria (28.9%, n=160). Co-infection with two different pathogens was also commonplace, comprising a total of 71 (12.8%) cases (Fig. 2A). Among the pathogens identified, most were common respiratory pathogens such as influenza viruses, rhinoviruses, *Pseudomonas aeruginosa* and *Haemophilus influenzae* (Fig. 2B). In addition, we identified a number of unconventional respiratory pathogens that are often not included in respiratory pathogen screening panels but known to cause severe infections in respiratory tract or lungs, including enterovirus D68 and *Chlamydia psittaci* (see below).

Finally, none of the pathogens described here appeared in the blank controls. Since the blank control samples were generated using the same procedures for RNA extraction, library preparation and sequencing as the experimental groups, these results effectively exclude the possibility that the pathogens described above were of contaminant origin.

Viruses

RNA viral pathogens exhibited both great diversity (15 species) and abundance (up to 52% of total RNA) in the BALF samples examined here. The most frequently detected RNA viruses were human rhinoviruses A-C (HRV, n=55), followed by influenza A virus (IAV, n=29), human

parainfluenza virus type 3 (HPIV3, n=20), influenza B virus (IBV, n=8), and human metapneumovirus (HPMV, n=7) (Fig. 2B). While the majority were found throughout the study period, a few had more specific time-scales (Fig. 2C). For example, influenza B viruses were mostly identified in 2017, whereas enterovirus D (ENV-D, n=6) was only detected in the summer of 2016. In addition, we identified all four types of common cold associated coronaviruses - OC43 (n=4), HKU1 (n=4), 229E (n=6) and NL63 (n=1) - all of which had a relatively low prevalence in our cohort. Importantly, none of the libraries contained any hit to SARS-CoV or SARS-CoV-2, these results were confirmed by both read mapping and blasting against corresponding virus genomes.

In comparison to RNA viruses, the DNA viruses identified were limited in diversity and abundance. All three major types of human herpesviruses were identified - HSV1 (n=3), CMV (n=3), and EBV (n=2) - although with low abundance in all cases (up to 30 RPM or 0.003%) (Fig. 2C). Another common DNA virus that causes respiratory disease – adenovirus - was also identified in several cases, although at abundance levels lower than the 1 RPM threshold such that it was considered a 'negative' result in this context.

Bacteria and fungi

The most common bacterial pathogens identified included *Pseudomonas aeruginosa* (n=31), *Haemophilus influenzae* (n=19), *Staphylococcus aureus* (n=18), and *Mycoplasma pneumoniae* (n=11), all of which are common respiratory pathogens. *Acinetobacter* bacteria were also prevalent in our cohort, including *Acinetobacter baumannii* (n=5) and *A. pittii* (n=1), both of which are commonly associated with hospital-acquired infections. Other important respiratory pathogens, such as *Mycobacterium tuberculosis* (n=2), *Legionella peumophila* (n=2),

Streptococcus pneumoniae (n=2), *Klebsiella pneumoniae* (n=3) and *Moraxella catarrhalis* (n=3), were also detected, although at a relatively low prevalence. Of particular interest was the identification of a single case of *Chlamydia psittaci* – a potentially bird-associated zoonotic pathogen – present in the BLAF at relatively high abundance (6396 RPM). Conversely, all the fungal pathogens identified here – *Candida albican* (n=3), *C. tropicalis* (n=1), and *Aspergillus spp.* (n=1) – were common pathogens known to cause respiratory infections.

qPCR confirmation of pathogen presence and abundance

For each of the pathogens identified here, we performed a qRT-PCR assay to confirm their presence and validate their abundance level as measured using our meta-transcriptomic approach. Strikingly, strong correlations were observed between the abundance measured by qPCR (i.e., CT value) and those estimated by read count after log 2 conversion (-0.8 < Pearson's R < -1, Fig. 3). Hence, the quantification by the two methods is strongly comparable.

In-depth phylogenetic characterization of pathogens

Although no novel viral pathogens were identified in this study, those viruses detected were characterized by substantial phylogenetic diversity, reflected in the presence of multiple viral lineages that highlight their complex epidemiological history in Wuhan (Fig. 4). We identified more than 14 genomic types of rhinovirus A, 7 of rhinovirus B, and 4 of rhinovirus C. A similar pattern of the co-circulation of multiple viral lineages was observed in other viruses. For example, the influenza A viruses identified in this study can be divided into the H1N1 and H3N2 subtypes, each containing multiple lineages that clustered with viruses sampled globally and reflecting the highly connected nature of Wuhan (Fig. 4). In this context it was striking that in the case of several coronaviruses - OC43, HKU1 and 229E - the Wuhan sequences grouped directly

153 with those identified from the United States (Fig. 4), although this may reflect limited sampling.

154 **Pathogen presence and abundance in diseased and healthy individuals**

155 Our meta-transcriptomic analysis revealed that many RNA viruses and bacteria detected were
156 present at extremely high abundance levels (>1%, and up to 52% of total RNA) and hence likely
157 indicative of acute disease. This was particularly true of eight species of RNA viruses – EV-D68,
158 the influenza viruses, HRV, HPIV3, 229E – as well as two species of bacteria (*Haemophilus*
159 *influenzae* and *Pseudomonas aeruginosa*) (Fig. 5). Together, these comprise a total of 54 cases
160 (13.2% of total diseases cases).

161 In marked comparison, high levels of abundance were never observed in the healthy control
162 group (Fig. 5). The highest abundance was 1302 RPM (0.013% of total RNA) in this group was
163 for *Haemophilus influenzae*. Indeed, for most of the pathogens, the abundance level in healthy
164 group was either undetectable or well below that observed in the diseased group, with the
165 exception of Human parainfluenza virus type 1 (HPIV) and *Mycoplasma orale* for which the
166 abundance levels were higher in the control group, although the sample size for both pathogens
167 was relatively small. Conversely, commensal microbes, particularly *Escherichia coli*, exhibited
168 similar abundance levels in the disease and control groups.

169

170 **Discussion**

171 Our study provides a critical snapshot of the respiratory pathogens present in Wuhan prior to the
172 emergence of SARS-CoV-2. Our unbiased metagenomic survey in patients presenting with
173 pneumonia or acute respiratory infections provides strong evidence that SARS-CoV-2 or any
174 related SARS-like viruses were absent in Wuhan approximately two years prior to the onset of

pandemic, although a number of common cold coronaviruses (HKU1, OC43, 229E, and NL63) were commonly detected in our cohort. Indeed, the earliest COVID-19 case, identified by qRT-PCR or next-generation sequencing-based assays performed at designated authoritative laboratories, can currently only be traced back to early December 2019 in Wuhan[9]. In addition, a retrospective survey of 640 throat swabs from patients with influenza-like illness in Wuhan from the period between 6 October 2019 and 21 January 2020 did not find any evidence of SARS-CoV-2 infection prior to January 2020[10], such that the ultimate origin of SARS-CoV-2 remains elusive[11].

The data presented provide a comprehensive overview of the infectome associated with pneumonia or acute respiratory infections in Wuhan, which is clearly more diverse and complex than described using previous surveys based on targeted PCR or qPCR approach alone[12]. In light of our observations, we can divide the respiratory infectome into three categories based on epidemiological characteristics: (i) the "core" infectome that is commonly found in patients with respiratory infection and is expected to occur each year globally, (ii) an "emerging" infectome that occurs during outbreaks but are not typically found in the geographic regions under investigation, and (iii) the sporadic infection or zoonotic infectome of new or rare pathogens.

The core infectome comprised of a wide range of common respiratory or systemic pathogens that are subject to frequent screening in hospitals. These include influenza viruses, HMPV, RSV, *Moraxella catarrhalis*, *Acinetobacter spp.*, *Klebsiella pneumoniae*, *Mycoplasma spp.*, *Haemophilus influenzae*, *Pseudomonas aeruginosa*, *Staphylococcus aureus* and *Streptococcus pneumoniae*. However, the remaining pathogens identified here, including rhinoviruses, parainfluenza viruses, coronaviruses, and herpesviruses, have often received far less attention

197 from clinicians and are sometimes ignored entirely, most likely due to the lack of association
198 with severe disease in adults[13-15]. Nevertheless, our results showed these "neglected" respiratory
199 viruses had high diversity, abundance and prevalence in the cohort of pneumonia or acute
200 respiratory patients studied here in comparison to healthy controls, such that their role as agents
201 of disease should not be underestimated. One scenario is that they represent opportunistic
202 pathogens that take advantage of weakened immunity, such as herpesviruses associated acute
203 respiratory distress (ARDS)[16]. It is also possible that their pathogenic effects have yet to be
204 identified and may extend to disease manifestations beyond respiratory infections. For example,
205 deep sequencing of a brain biopsy sample suggesting that the OC43 coronavirus may be
206 associated with fatal encephalitis in humans[17].

207 We also identified a potential "emerging" infectome, in this case comprising a single virus - EV-
208 D68 - that may represent a regional or national outbreak of an unconventional respiratory
209 pathogens. The prevalence of EV-D68 remained low from its discovery in 1967 until 2014[18],
210 when a major outbreak started in the United States and spread to more than 20 countries[19],
211 causing severe respiratory illness with potential neurological manifestations such as acute flaccid
212 paralysis in children[20, 21]. In China, EV-D68 was only sporadically reported[22, 23], although
213 serological surveys suggest a much wider prevalence for both children and adults since 2009[24].
214 We identified six EV-D68 cases, all adults that occurred within a relatively narrow time window
215 between June and December 2016. Phylogenetic analysis revealed that the sequences of these
216 viruses were closely related to each other and to other Chinese strains from the same period (Fig.
217 4), suggesting that it may be a part of a larger outbreak in China. The EV-D68 cases identified
218 here showed moderate to severe respiratory symptoms, although viral abundance was generally
219 very high, with four of six cases showing >106 RPM (i.e., >10% of total RNA) in the BALF

sample. This highlights the active replication and massive proliferation of viruses within the respiratory system of these patients.

Finally, our zoonotic infectome also comprised a single pathogen, *Chlamydia psittaci*, that is associated with avian species but causes occasional outbreaks in domestic animals (i.e., pigs, cattles, and sheep) and humans[25]. In humans, *C. psittaci* infections often starts with influenza-like symptoms, but can develop into serious lung infections and even death[26]. The single case of *C. psittaci* identified here was at relatively high abundance level (6396 RPM) and caused a relatively severe disease, with the patient experiencing expiratory dyspnea, severe pneumonia and pleural effusion, and was subsequently transfer to intensive care unit (ICU) for further treatment. Since the patient had no travel history for a month prior to illness, this discovery underlines the risk of local exposure to this bacterium.

Our phylogenetic analysis of the metagenomic data generated revealed extensive intra-specific diversity in each virus species identified highlights the complex epidemiological history of these pathogens. Indeed, the influenza A viruses (H1N1 and H3N2), influenza B virus, HPIV3 and HCoV-OC43 discovered here all comprised multiple lineages (Fig. 4), suggesting these viruses were introduced from diverse sources. Since some of the viruses were closely related to those circulating in other countries it is possible that they represent overseas importations: this is not surprising given that Wuhan is a major domestic travel hub and well linked internationally. Indeed, the rapid and widespread transmission of SARS-CoV-2 between Wuhan and other major cities globally was key to seeding the global pandemic of COVID-19, with early cases in a number of localities all linked to travel from Wuhan[26-29].

The methodology used here served as an unbiased investigation of potential emerging pathogens.

Meta-transcriptomics has several advantages over traditional diagnostic approaches based on serology or PCR: it targets all types of micro-organisms simultaneously, identifies potential pathogens without *a priori* knowledge of what micro-organisms might be present, reveals the information (RNA) expressed by the pathogen during infection that is central to agent identification and studies of disease association. With the popularization of next-generation sequencing platform in major hospitals, the approach outlined here can be easily integrated into diagnostic practice with much greater speed and significantly more information output than traditional technologies, providing a broad-scale understanding of infectious disease in general.

Material and Methods

Ethics statement

The sampling and experimental procedures for this study were reviewed by the ethics committees of the Central Hospital of Wuhan and the National Institute for Communicable Disease Control and Prevention, Chinese Center for Disease Control and Prevention. Written informed consents were taken from all patients and volunteers recruited in this study. In addition, for child patients, written consents were also obtained from their parents or guardians. Physicians were informed of results of the pathogen discovery exercise as soon as the meta-transcriptomic results were obtained.

Sample collection from patients and controls.

More than 1000 patients were recruited from the Central Hospital of Wuhan between 2016 and 2017. The target clinical conditions were community-acquired pneumonia and acute respiratory infections based on the initial diagnosis made by clinicians. All patients were hospitalized and

subject to bronchoalveolar lavage fluid (BALF) collection required by the initial diagnosis for pneumonia or acute respiratory distress syndrome and independent of this study. The BALF sample was divided into two parts for the clinical laboratory test and this study, respectively. Of the BALF samples collected, 409 were subjected to meta-transcriptomic analysis based on their condition and the time between hospitalization and sample collection. No diagnostic information was provided prior to sample selection. To establish a healthy control group, 5ml-10ml of BALF samples were also taken from 32 volunteers without respiratory symptoms. We also included 10 blank controls where only RNase free water was used for nucleic acid extraction and library construction, although only four of these produced viable RNA sequencing results.

Meta-transcriptomic pathogen discovery pipeline

We followed a standard protocol for meta-transcriptomics analysis for each BALF sample. Total RNA was first extracted from 200–300ul of each sample using the RNeasy Plus Universal Kit (Qiagen, USA) according to the manufacturer's instructions. From the extracted RNA, we performed human rRNA removal and low concentration library construction procedures with the Trio RNA-Seq kit (NuGEN Technologies, USA). The libraries were then subjected to 150bp pair-end sequencing on an Illumina Hiseq 4000 platform at Novagene (Beijing), with target output of 10G base pairs per library. For each of the sequencing results generated, we removed adaptor sequences, non-complex reads, as well as duplicated reads using the BBmap software package. Human and ribosomal RNA (rRNA) reads were subsequently removed by mapping the de-duplicated reads against the human reference genome (GRCh38/hg38) and the comprehensive rRNA sequence collection downloaded from SILVA database[30].

The remaining sequencing reads were subject to a pathogen discovery pipeline. For virus

identification, the reads were directly compared against reference virus databases using the blastn program and against the non-redundant protein (nr) database using diamond blastx[31], with an e-value threshold set at 1E-10 and 1E-5 for blastn and diamond blastx analyses, respectively. Viral abundance was summarized from both analyses, calculated using the relation: total viral reads/total non-redundant reads* 1 million (i.e., reads per million, RPM). To identify highly divergent virus genomes, reads were assembled using megahit[32] into contigs before comparison against the nt and nr databases. Those reads with <90% amino acid similarity to known viruses were treated as potential novel virus species. For bacterial and fungi identification, we first used MetaPhlAn2[33] to identify potential species in both groups. Relevant background bacterial and fungal genomes were subsequently downloaded from NCBI/GenBank and used as a template for read mapping. Based on the mapping results for each case, we generated relevant contigs for blastn analyses against the non-redundant nucleotide (nt) database to determine taxonomy to the species level. The abundance level of bacterial and fungi pathogens was also calculated in the form of RPM based on genome and mitochondrial genome read counts, respectively.

A microbe was considered as "positive" within a specific sample if its abundance level was greater than 1 RPM. To prevent false positives resulting from index hopping, we used a threshold of 0.1% for viruses present in the same sequencing lane: that is, if the libraries contain less than 0.1% of the most abundant library it is treated as "negative".

Confirmatory testing by conventional methods

For pathogen positive samples, the same RNA used for meta-transcriptomics analysis was also subject to a qRT-PCR assay with primers sets designed for a specific or related group of pathogens (Table S1). RNA was first reverse transcribed by SuperScript™ III First-Strand

Synthesis SuperMix for qRT-PCR (Invitrogen, California), and then amplified by TaqPath™ ProAmp™ Master Mix (Applied Biosystems, California). A cycle threshold (CT) value of 38 and above was treated as negative.

Pathogen genomic analyses

For viruses at high abundance levels (i.e., > 1000 RPM), complete genomes were assembled using Megahit and confirmed by mapping reads against the assembled contigs. These genomes were then aligned with related reference virus sequences downloaded from NCBI/GenBank using MAFFT program[34]. Ambiguously aligned regions were removed using the Trimal program[35]. Phylogenetic trees were estimated using the maximum likelihood approach implemented in PhyML[36], employing the GTR model of nucleotide substitution and SPR branch swapping. The support for each node in the tree was estimated with an approximate likelihood ratio test (aLRT) with the Shimodaira-Hasegawa-like procedures.

Data availability

All non-human reads have been deposited in the SRA databases under the project accession PRJNA699976. Relevant consensus virus genome sequences have been deposited in NCBI/GenBank under the accessions MW567157- MW567162, MW570805- MW570808, MW571087- MW571107, MW587035- MW587095.

Acknowledgments

This study was supported by the National Natural Science Foundation of China 32041004 (YZZ), 31930001 (YZZ), 81861138003 (YZZ) and 81672057 (YZZ). E.C.H. is supported by an

ARC Australian Laureate Fellowship (FL170100022).

Author Contributions

Yong-Zhen Zhang conceived and designed the study. Su Zhao, Yi Hu, Wen Yin, Fang Ni, Hong-Ling Hu, Shuang Geng, and Li Tan performed the clinical work and sample collection. Bin Yu, Jun-Hua Tian and Ying Peng performed epidemiological investigation and sample collection. Mang Shi, Wei-Chen Wu, Yan-Mei Chen, Wen Wang, and Zhi-Gang Song performed the experimental works. Mand Shi, Edwards C Holmes and Yong-Zhen Zhang analysed the data. Mang Shi, Edward C Holmes and Yong-Zhen Zhang wrote the paper with input from all authors.

Competing Interests

All other authors declare no competing interests.

References

1. Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, Hu Y, Tao ZW, Tian JH, Pei YY, Yuan ML, Zhang YL, Dai FH, Liu Y, Wang QM, Zheng JJ, Xu L, Holmes EC, Zhang YZ.A new coronavirus associated with human respiratory disease in China. *Nature* 2020; 579:265-9.
2. Z Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, Si HR, Zhu Y, Li B, Huang CL, Chen HD, Chen J, Luo Y, Guo H, Jiang RD, Liu MQ, Chen Y, Shen XR, Wang X, Zheng XS, Zhao K, Chen QJ, Deng F, Liu LL, Yan B, Zhan FX, Wang YY, Xiao GF, Shi ZL. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* 2020; 579:270-3.
3. Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, Ren R, Leung KSM, Lau EHY, Wong JY, Xing X, Xiang N, Wu Y, Li C, Chen Q, Li D, Liu T, Zhao J, Liu M, Tu W, Chen C, Jin L, Yang R, Wang Q, Zhou S, Wang R, Liu H, Luo Y, Liu Y, Shao G, Li H, Tao Z, Yang Y, Deng Z, Liu B, Ma Z, Zhang Y, Shi G, Lam TTY, Wu JT, Gao GF, Cowling BJ, Yang B, Leung GM, Feng Z. Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. *N Engl J Med* 2020; 382:1199-207.
4. Zhang YZ, Chen YM, Wang W, Qin XC, Holmes EC. Expanding the RNA virosphere by unbiased metagenomics. *Annu Rev Virol*; 2019; 6:119-139.
5. Eden JS, Rose K, Ng J, Shi M, Wang Q, Sintchenko V, Holmes EC. Francisella tularensis ssp. holarctica in Ringtail Possums, Australia. *Emerg Infect Dis* 2017; 23:1198-201.
6. Shi M, Lin XD, Chen X, Tian JH, Chen LJ, Li K, Wang W, Eden JS, Shen JJ, Liu L, Holmes EC, Zhang YZ. The evolutionary history of vertebrate RNA viruses. *Nature* 2018; 556:197-202.
7. Shi M, Lin XD, Tian JH, Chen LJ, Chen X, Li CX, Qin XC, Li J, Cao JP, Eden JS, Buchmann J, Wang W, Xu JG, Holmes EC, Zhang YZ. Redefining the invertebrate RNA virosphere. *Nature* 2016; 540:539-43.
8. Li CX, Li W, Zhou J, Zhang B, Feng Y, Xu CP, Lu YY, Holmes EC, Shi M. High resolution metagenomic characterization of complex infectomes in paediatric acute respiratory

infection. *Sci Rep* 2020; 10:3963.

9. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, Zhang L, Fan G, Xu J, Gu X, Cheng Z, Yu T, Xia J, Wei Y, Wu W, Xie X, Yin W, Li H, Liu M, Xiao Y, Gao H, Guo L, Xie J, Wang G, Jiang R, Gao Z, Jin Q, Wang J, Cao B. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* 2020; 395:497-506.
10. Kong WH, Li Y, Peng MW, Kong DG, Yang XB, Wang L, Liu MQ. SARS-CoV-2 detection in patients with influenza-like illness. *Nat Microbiol* 2020; 5:675-8.
11. Hu B, Guo H, Zhou P, Shi ZL. Characteristics of SARS-CoV-2 and COVID-19. *Nat Rev Microbiol* 2020.
12. Wu Z, Li Y, Gu J, Zheng H, Tong Y, Wu Q. Detection of viruses and atypical bacteria associated with acute respiratory infection of children in Hubei, China. *Respirology* 2014; 19:218-24.
13. Jacobs SE, Lamson DM, St George K, Walsh TJ. Human rhinoviruses. *Clin Microbiol Rev* 2013;26:135-62.
14. Henrickson KJ. Parainfluenza viruses. *Clin Microbiol Rev* 2003;16:242-64.
15. van der Hoek L. Human coronaviruses: what do they cause? *Antivir Ther* 2007;12:651-8.
16. Bonizzoli M, Arvia R, di Valvasone S, Liotta F, Zakrzewska K, Azzi A, Peris A. Human herpesviruses respiratory infections in patients with acute respiratory distress (ARDS). *Med Microbiol Immunol* 2016; 205:371-9.
17. Morfopoulou S, Brown JR, Davies EG, Anderson G, Virasami A, Qasim W, Chong WK, Hubank M, Plagnol V, Desforges M, Jacques TS, Talbot PJ, Breuer J. Human coronavirus OC43 associated with fatal encephalitis. *N Engl J Med* 2016; 375:497-8.
18. Schieble JH, Fox VL, Lennette EH. A probable new human picornavirus associated with respiratory diseases. *Am J Epidemiol* 1967; 85:297-310.
19. Holm-Hansen CC, Midgley SE, Fischer TK. Global emergence of enterovirus D68: a systematic review. *Lancet Infect Dis* 2016; 16:e64-e75.
20. Midgley CM, Watson JT, Nix WA, Curns AT, Rogers SL, Brown BA, Conover C,

Dominguez SR, Feikin DR, Gray S, Hassan F, Hoferka S, Jackson MA, Johnson D, Leshem E, Miller L, Nichols JB, Nyquist AC, Obringer E, Patel A, Patel M, Rha B, Schneider E, Schuster JE, Selvarangan R, Seward JF, Turabelidze G, Oberste MS, Pallansch MA, Gerber SI; EV-D68 Working Group. Severe respiratory illness associated with a nationwide outbreak of enterovirus D68 in the USA (2014): a descriptive epidemiological investigation. *Lancet Respir Med* 2015; 3:879-87.

21. Greninger AL, Naccache SN, Messacar K, Clayton A, Yu G, Somasekar S, Federman S, Stryke D, Anderson C, Yagi S, Messenger S, Wadford D, Xia D, Watt JP, Van Haren K, Dominguez SR, Glaser C, Aldrovandi G, Chiu CY. A novel outbreak enterovirus D68 strain associated with acute flaccid myelitis cases in the USA (2012-14): a retrospective cohort study. *Lancet Infect Dis* 2015;15:671-82.
22. Xiang Z, Gonzalez R, Wang Z, Ren L, Xiao Y, Li J, Li Y, Vernet G, Paranhos-Baccalà G, Jin Q, Wang J. Coxsackievirus A21, enterovirus 68, and acute respiratory tract infection, China. *Emerg Infect Dis* 2012;18:821-4.
23. Xiao Q, Ren L, Zheng S, Wang L, Xie X, Deng Y, Zhao Y, Zhao X, Luo Z, Fu Z, Huang A, Liu E. Prevalence and molecular characterizations of enterovirus D68 among children with acute respiratory infection in China between 2012 and 2014. *Sci Rep* 2015; 5:16639.
24. iang Z, Li L, Ren L, Guo L, Xie Z, Liu C, Li T, Luo M, Paranhos-Baccalà G, Xu W, Wang J. Seroepidemiology of enterovirus D68 infection in China. *Emerg Microbes Infect* 2017; 6:e32.
25. Knittler MR, Sachse K. Chlamydia psittaci: update on an underestimated zoonotic agent. *Pathog Dis* 2015; 73:1-15.
26. Watanabe M. The COVID-19 Pandemic in Japan. *Surg Today* 2020; 50:787-93.
27. Lescure FX, Bouadma L, Nguyen D, Parisey M, Wicky PH, Behillil S, Gaymard A, Bouscambert-Duchamp M, Donati F, Le Hingrat Q, Enouf V, Houhou-Fidouh N, Valette M, Mailles A, Lucet JC, Mentre F, Duval X, Descamps D, Malvy D, Timsit JF, Lina B, van-der-Werf S, Yazdanpanah Y.. Clinical and virological data of the first cases of COVID-19 in Europe: a case series. *Lancet Infect Dis* 2020; 20:697-706.
28. Wei WE, Li Z, Chiew CJ, Yong SE, Toh MP, Lee VJ. Presymptomatic transmission of

426 SARS-CoV-2 - Singapore, January 23-March 16, 2020. *MMWR Morb Mortal Wkly Rep*
427 2020; 69:411-5.
428 29. Song JY, Yun JG, Noh JY, Cheong HJ, Kim WJ. Covid-19 in South Korea - Challenges of
429 subclinical manifestations. *N Engl J Med* 2020; 382:1858-9.
430 30. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. The
431 SILVA ribosomal RNA gene database project: improved data processing and web-based
432 tools. *Nucleic Acids Res* 2013; 41:D590-6.
433 31. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat*
434 *Methods* 2015; 12:59-60.
435 32. Li D, Liu CM, Luo R, Sadakane K, Lam TW. MEGAHIT: an ultra-fast single-node solution
436 for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*
437 2015; 31:1674-6.
438 33. Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, Tett A, Huttenhower
439 C, Segata N.. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat Methods*
440 2015; 12:902-3.
441 34. Katoh K, Toh H. Recent developments in the MAFFT multiple sequence alignment program.
442 *Brief Bioinform* 2008; 9:286-98.
443 35. Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. trimAl: a tool for automated alignment
444 trimming in large-scale phylogenetic analyses. *Bioinformatics* 2009; 25:1972-3.
445 36. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms
446 and methods to estimate maximum-likelihood phylogenies: assessing the performance of
447 PhyML 3.0. *Syst Biol* 2010; 59:307-21.
448

449 **Figure legends**

450 **Fig. 1. Patient recruitment**. (A) Sampling frequency and intensity during the study period. (B)
451 Male-to-female ratio of the recruited patients. (C) Age structure of all patients involved in this
452 study depicted using violin plots. (D) Type and frequencies of pre-existing conditions. (E) The
453 number of severe, non-severe and control cases.

454

455 **Fig. 2. Prevalence and abundance of viral, bacterial and fungal pathogens in the 408 cases**
456 **examined in this study**. (A) Proportion of cases infected with RNA virus, DNA virus, bacteria,
457 fungus and with mixed infections. (B) Prevalence of each pathogen, ordered by the number of
458 cases. (C) Heat map showing the prevalence and abundance of pathogens in diseased and control
459 samples. *Escherichia coli* is not regarded as pathogen but is shown here as an example of non-
460 pathogen contamination. The samples (x-axis) are divided into "Patient" and "Control" groups,
461 each ordered chronically. The pathogens (y-axis) are divided into four categories: RNA viruses,
462 DNA viruses, bacteria, and fungi.

463

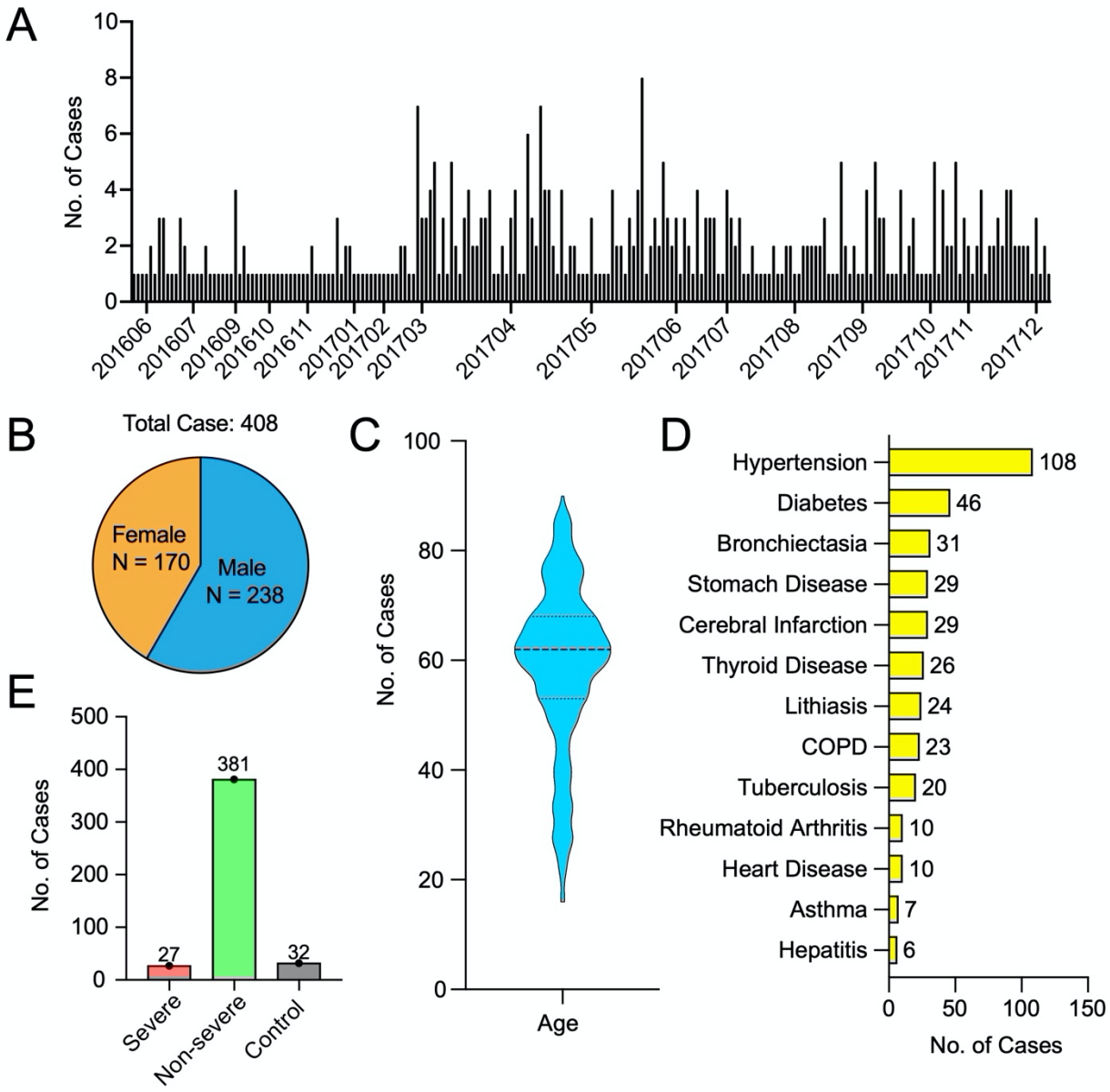
464 **Fig. 3. Comparisons of pathogen abundance measured by qPCR and meta-transcriptomics**
465 **approaches**. Abundance by qPCR methods is measured by cycle threshold, or CT value, while
466 those by meta-transcriptomics are measured by "log2(number of reads)". The comparisons are
467 performed on four most abundant pathogens: influenza A virus, influenza B virus, human
468 rhinoviruses, and human parainfluenza virus 3. Pearson's correlation coefficient between CT
469 values and log2(number of reads) is estimated for each pathogen.

470

471 **Fig. 4. Evolutionary relationships of RNA virus pathogens identified in this study**.

Phylogenetic trees were estimated using the maximum likelihood method implemented in PhyML. Sequences identified from this study were marked with red solid circle. For larger trees, we only show lineages or sub-lineages which contain sequences identified in this study.

Fig. 5. Comparison of prevalence levels between healthy and control groups. Boxplots for patient and control groups for each pathogen identified in this study, including RNA virus (green), DNA virus (yellow), bacteria (orange) and fungus (blue). For clarity, only non-zero abundance levels are reported.



480

481 **Fig. 1. Patient recruitment**. (A) Sampling frequency and intensity during the study period. (B)
482 Male-to-female ratio of the recruited patients. (C) Age structure of all patients involved in this
483 study depicted using violin plots. (D) Type and frequencies of pre-existing conditions. (E) The
484 number of severe, non-severe and control cases.

485

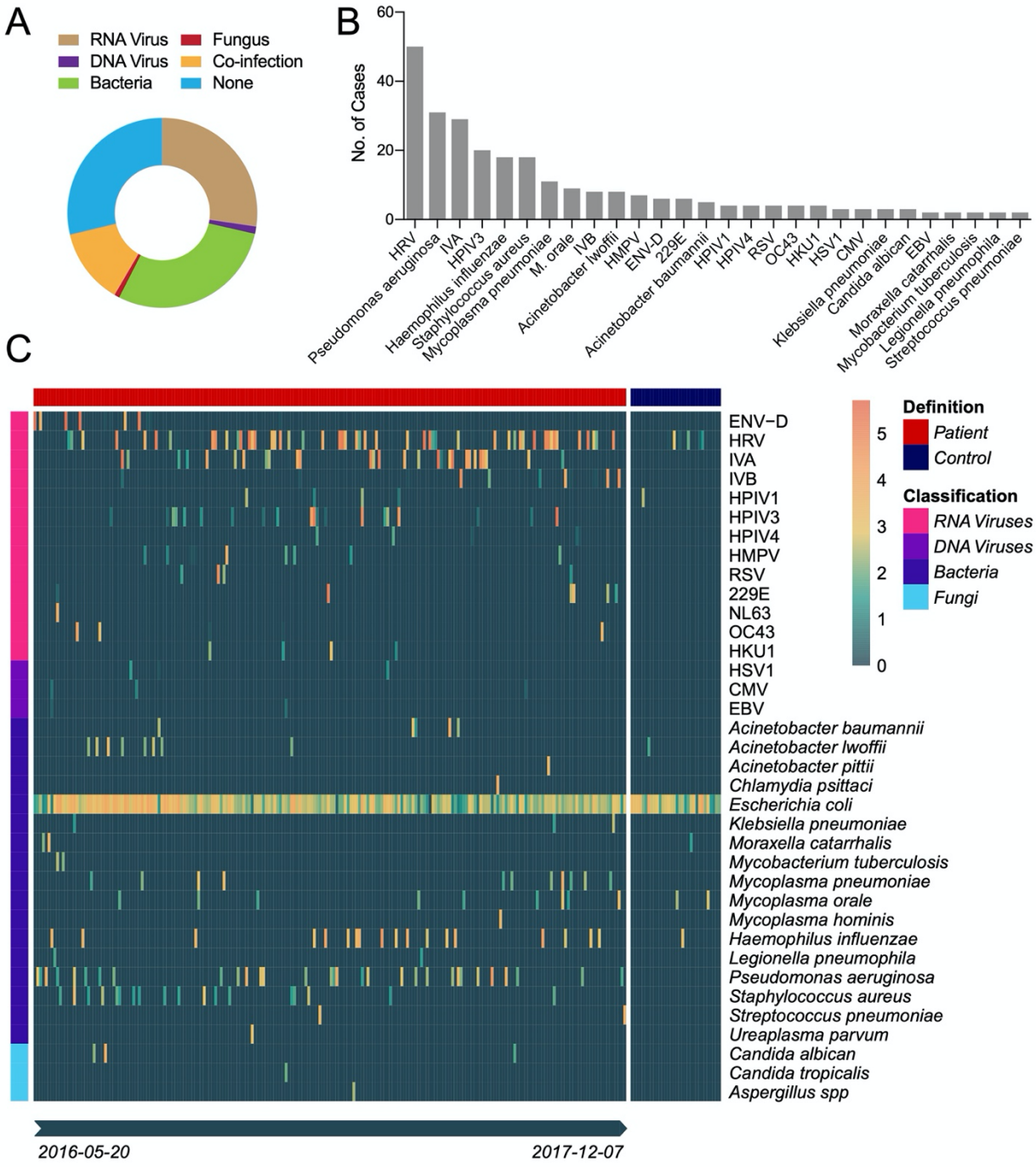


Fig. 2. Prevalence and abundance of viral, bacterial and fungal pathogens in the 408 cases examined in this study. (A) Proportion of cases infected with RNA virus, DNA virus, bacteria, fungus and with mixed infections. (B) Prevalence of each pathogen, ordered by the number of cases. (C) Heat map showing the prevalence and abundance of pathogens in diseased and control

491 samples. *Escherichia coli* is not regarded as pathogen but is shown here as an example of non-
492 pathogen contamination. The samples (x-axis) are divided into "Patient" and "Control" groups,
493 each ordered chronically. The pathogens (y-axis) are divided into four categories: RNA viruses,
494 DNA viruses, bacteria, and fungi.
495

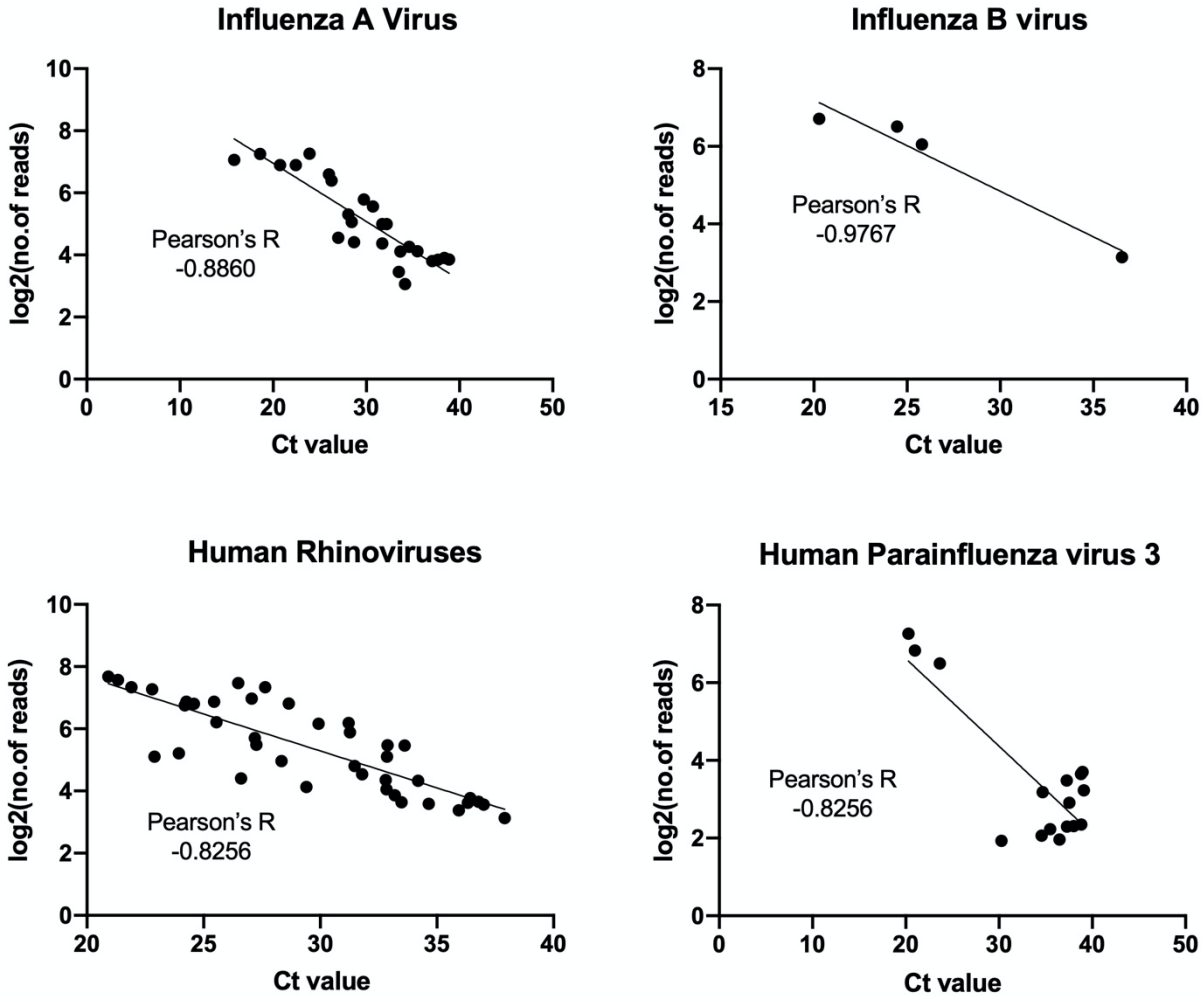
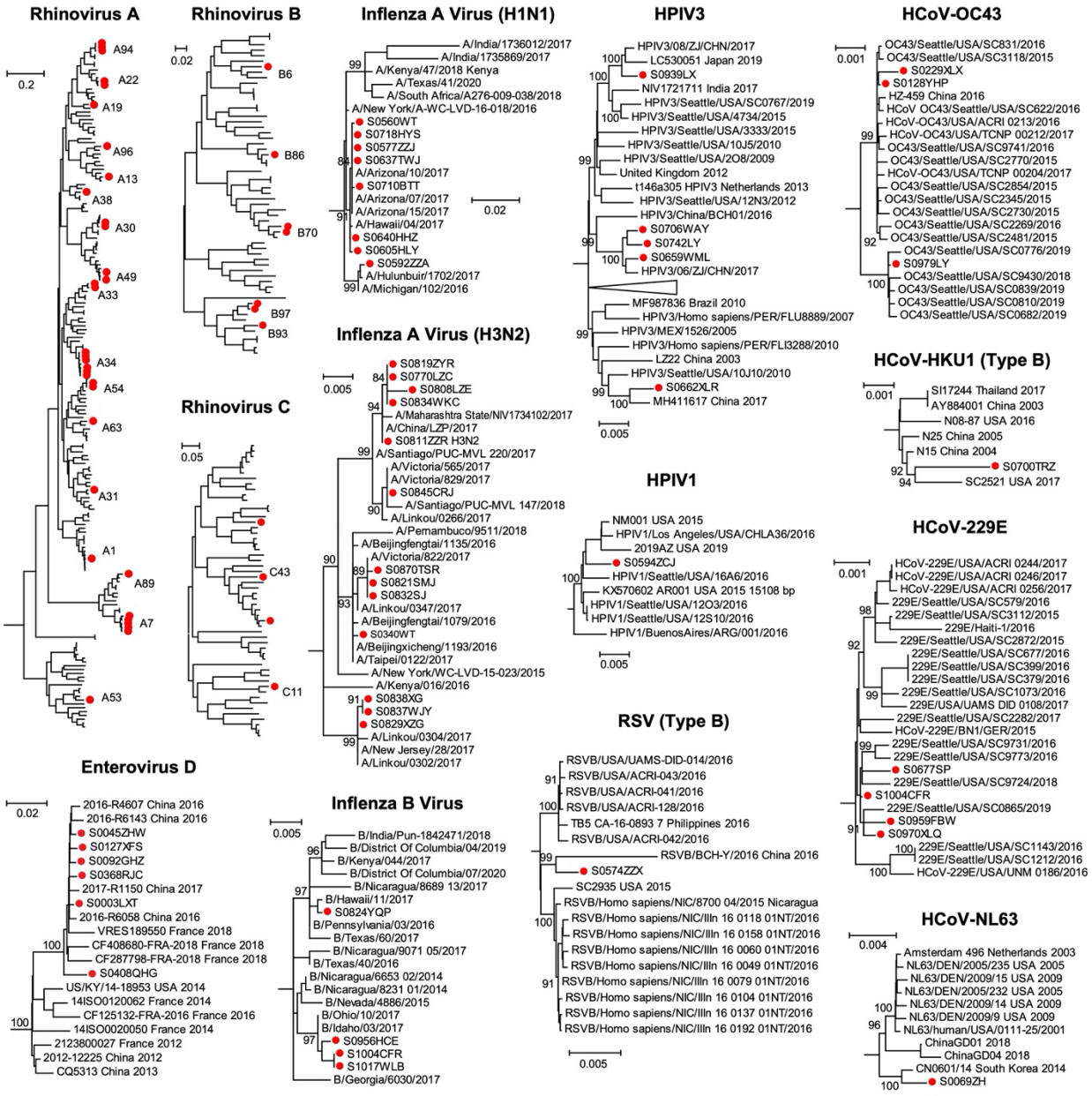


Fig. 3. Comparisons of pathogen abundance measured by qPCR and meta-transcriptomics approaches. Abundance by qPCR methods is measured by cycle threshold, or CT value, while those by meta-transcriptomics are measured by "log2(number of reads)". The comparisons are performed on four most abundant pathogens: influenza A virus, influenza B virus, human rhinoviruses, and human parainfluenza virus 3. Pearson's correlation coefficient between CT values and log2(number of reads) is estimated for each pathogen.



504

505 **Fig. 4. Evolutionary relationships of RNA virus pathogens identified in this study**.

506 Phylogenetic trees were estimated using the maximum likelihood method implemented in

507 PhyML. Sequences identified from this study were marked with red solid circle. For larger trees,

508 we only show lineages or sub-lineages which contain sequences identified in this study.

509

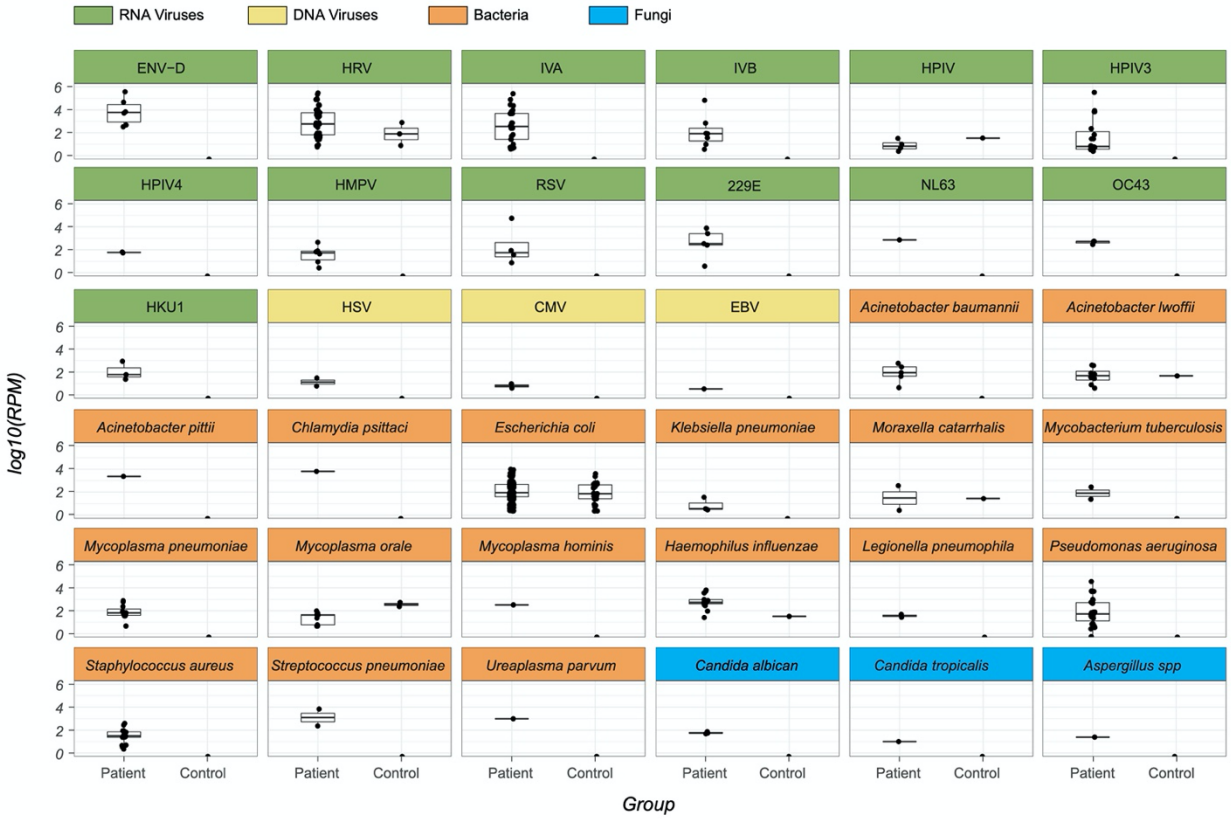


Fig. 5. Comparison of prevalence levels between healthy and control groups. Boxplots for patient and control groups for each pathogen identified in this study, including RNA virus (green), DNA virus (yellow), bacteria (orange) and fungus (blue). For clarity, only non-zero abundance levels are reported.