Check for updates

RESEARCH ARTICLE

## REVISED UK circulating strains of human parainfluenza 3: an amplicon based next generation sequencing method and phylogenetic analysis [version 2; referees: 2 approved]

Anna Smielewska [iD] [1,2], Edward Emmott [iD] [1,3], Kyriaki Ranellou [iD] [1,2], Ashley Popay[4], Ian Goodfellow [iD] [1], Hamid Jalal [iD] [2]

[1]Department of Pathology, University of Cambridge Addenbrooke's Hospital Cambridge, Cambridge, Cambridgeshire, CB20QQ, UK
[2]Cambridge University Hospitals NHS Foundation Trust Laboratory, Public Health England, Cambridge, Cambridgeshire, CB20QQ, UK
[3]Department of Bioengineering, Northeastern University, Boston, MA, 02115-5000, USA
[4]Eastern Field Epidemiology Unit, Institute of Public Health, Public Health England, Cambridge, Cambridgeshire, CB20SR, UK

## Abstract

**Background:** Human parainfluenza viruses type 3 (HPIV3) are a prominent cause of respiratory infection with a significant impact in both pediatric and transplant patient cohorts. Currently there is a paucity of whole genome sequence data that would allow for detailed epidemiological and phylogenetic analysis of circulating strains in the UK. Although it is known that HPIV3 peaks annually in the UK, to date there are no whole genome sequences of HPIV3 UK strains available.

**Methods:** Clinical strains were obtained from HPIV3 positive respiratory patient samples collected between 2011 and 2015. These were then amplified using an amplicon based method, sequenced on the Illumina platform and assembled using a new robust bioinformatics pipeline. Phylogenetic analysis was carried out in the context of other epidemiological studies and whole genome sequence data currently available with stringent exclusion of significantly culture-adapted strains of HPIV3.

**Results:** In the current paper we have presented twenty full genome sequences of UK circulating strains of HPIV3 and a detailed phylogenetic analysis thereof. We have analysed the variability along the HPIV3 genome and identified a short hypervariable region in the non-coding segment between the M (matrix) and F (fusion) genes. The epidemiological classifications obtained by using this region and whole genome data were then compared and found to be identical.

**Conclusions:** The majority of HPIV3 strains were observed at different geographical locations and with a wide temporal spread, reflecting the global distribution of HPIV3. Consistent with previous data, a particular subcluster or strain was not identified as specific to the UK, suggesting that a number of genetically diverse strains circulate at any one time. A small hypervariable region in the HPIV3 genome was identified and it was shown that, in the absence of full genome data, this region could be used for epidemiological surveillance of HPIV3.

## Open Peer Review

**Referee Status:** ✔ ✔

|  | Invited Referees | |
|---|---|---|
|  | **1** | **2** |
| REVISED **version 2** published 26 Nov 2018 | ✔ report | ✔ |
| **version 1** published 19 Sep 2018 | ? report | ✔ report |

1  **Hirokazu Kimura**, Gunma Paz University, Japan
   Gunma Paz University, Japan

2  **Julian Wei-Tze Tang**, University Hospitals of Leicester NHS Trust, UK
   University of Leicester, UK

## Discuss this article

Comments (0)

## Keywords

human parainfluenza 3, phylogenetics, epidemiology, circulating strains

**Corresponding author:** Anna Smielewska (aas31@cam.ac.uk)

**Author roles: Smielewska A**: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Software, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Emmott E**: Conceptualization, Methodology, Visualization, Writing – Review & Editing; **Ranellou K**: Methodology, Writing – Review & Editing; **Popay A**: Data Curation, Visualization, Writing – Review & Editing; **Goodfellow I**: Conceptualization, Funding Acquisition, Project Administration, Resources, Supervision, Writing – Review & Editing; **Jalal H**: Conceptualization, Funding Acquisition, Project Administration, Resources, Supervision, Writing – Review & Editing

## Introduction

Human parainfluenza viruses (HPIV) are members of the family Paramyxoviridae and are subdivided into four types, which fall into two genera *Rubulavirus (types 2 and 4 and Respirovirus (types 1 and 3).* Human parainfluenza 3 (HPIV3) (recently renamed human respirovirus 3) is a negative strand non-segmented RNA virus of 15462 nucleotides in length. It consists of a core containing the RNA bound to the nucleocapsid protein (NP), the phosphoprotein (P) and the large RNA polymerase (L) surrounded by an envelope composed of the matrix protein (M) and a lipid bilayer. The haemagluttinin neuraminidase (HN) protein and the fusion (F) protein are found on the envelope surface and facilitate the binding of HPIV3 to the sialic acid receptors of the target cell via the haemagluttinin component of HN, the fusion (F) with the cell and the release of new viral particles via the neuraminidase component of HN[1,2].

All four types of parainfluenza are significant causes of both upper and lower respiratory tract infections. Human parainfluenza 3 (HPIV3) has been identified as the most prevalent circulating parainfluenza serotype in the UK[3]. It is an important respiratory pathogen with a broad spectrum of presentations and a significant impact both in the paediatric and the immunocompromised cohorts. In the former it is responsible for up to 6.8% of all paediatric admissions for respiratory presentations[4]. In the immunocompromised population, the reported incidence of infection has varied between 5 and 12% with lower respiratory tract infections (LRTIs) and mortality of up to 75% being reported[5,6]. Transmission is by respiratory droplets and HPIV3 can persist up to 10 hours on non-absorbant surfaces[7]. Prolonged shedding in vulnerable patient groups leads to outbreaks and an increased burden on the health services[8,9]. Although a number of previous studies have looked at circulating HPIV3 strains throughout the world, there is currently no genetic data on the circulating HPIV3 strains in the UK.

Previous phylogenetic analysis of HPIV3 has been based on components of the genome rather than full-length genome data. To date there appears to be no unified phylogenetic classification of HPIV3. Recently the HN gene has been used to characterize emerging strains as well as tracing outbreaks[10–14]. Automatic barcode gap discovery (ABGD)[15] was used to separate HPIV3 into 3 clusters with currently circulating strains confined to one of these. In another study it was shown that the F gene is equally valid for HPIV3 phylogenetic classification[11]. The region directly preceding and overlapping with the start of the F genome has previously been identified as highly variable and has been historically used to trace outbreaks[8,9].

Additionally there is currently little experimental evidence linking HPIV3 HN and F phenotype with variations in the adaptive immune response[16,17] and currently no consensus on the connection between genotype and clinical pathogenesis of HPIV3. Therefore there is a clear need to rationalize the approach to the phylogenetic and epidemiological analysis of HPIV3.

To this end, in this study we have presented the genetic analysis of full genome sequences of twenty circulatingUK strains between the years 2011–2015. We have used this data, together with other full genome sequences available in the genebank to conduct a full genome phylogentic analysis of HPIV3. Although rapid metagenomic sequencing has been conducted in a small outbreak[18], given the relative expense of obtaining full genome data in a clinical setting we have identified a short hypervariable region in the HPIV3 genome and evaluated the reliability of using this segment for future phylogenetic analysis and potential epidemiological investigation.

## Methods
### Clinical samples
Clinical strains were obtained from HPIV3 positive respiratory patient samples collected between 2011 and 2015 by Public Health England (PHE) laboratory in a major teaching hospital. All identifiable information was removed prior to the study. Anonymous patient demographics such as age, sex, location, as well as date and type of the sample were retained where possible (ethics approval number 12/EE/0069).

All clinical strains were grown on PLC/PRF/5 human Alexander hepatoma cell line as described in a separate study[19] and underwent an additional passage for RNA harvesting. 43 samples were successfully grown and 20 clinical strains were selected for subsequent sequence analysis. Laboratory strain MK9 obtained from PHE cultures was used as a reference strain for sequencing pipeline validation.

### RNA extraction and amplification
Total RNA from samples was extracted using the GenElute Mammalian Total RNA Miniprep kit (RTN350, Sigma) according to the manufacturer's guidelines. Full genome amplification was achieved using a set of twelve primers producing twelve overlapping amplicons (Figure 1).

The Superscript III One-step RT-PCR System with Platinum Taq High Fidelity from Invitrogen (12574035 Invitrogen) was used for amplicon generation. The RT-PCR was carried out on the Eppendorf Mastercycler nexus GSX1. The RT step was performed at 50°C for 30 min. This was followed by a 2min

**A**

| Primer pair | | Primer sequence | Fragment (bases) |
|---|---|---|---|
| 1 | F | ACCAAACAAGAGAAGARACTTGTYTG | 1-520 |
| | R | TCCAGGTCACTTCCAAATATCCA | |
| 2 | F | CAGCYGGTGGAGCTATCATT | 166-2127 |
| | R | GAGCTGCTTCTTCTCCCAGG | |
| 3 | F | AAACAARGCAGTCAACCACC | 1569-2646 |
| | R | TCTGTTTGCCCCTTTGTGTC | |
| 4 | F | CGATTGCTCTCTTCTGTATCCTTCC | 2148-4212 |
| | R | TGGCATCRAACAGCATTCCT | |
| 5 | F | ACCCCACATTAGAGTTGCCA | 3860-5659 |
| | R | TCCTGCTGCTTCACAACCTA | |
| 6 | F | GGACACAAAYAAAGCAGTGC | 5527-7587 |
| | R | AGGAGTGCTAGAGARATGACT | |
| 7 | F | ACTCAGACTTGGTACCTGACT | 7500-9398 |
| | R | TGCACTTATATTCCATCGGCC | |
| 8 | F | AYGTGCTATCRAAATTAGCCTCA | 9046-11184 |
| | R | AGCYTGAGGAAGAATTCTCCCATCATA | |
| 9 | F | ATAGGYGTGAGGGTGACTGC | 10929-12433 |
| | R | TGTGAGGCTTCCATCCAAGA | |
| 10 | F | GGACTTGAAACRCCTGACCCA | 12111-14064 |
| | R | GGAAGAGCCTGTCCTTGTCT | |
| 11 | F | GAAGGTAGRGATCTCATTTGGGA | 13164-15067 |
| | R | CCTGTAAGYAATCGAGTCGAT | |
| 12 | F | TYGGAATCAACAGCACTAGTTG | 13975-15462 |
| | R | ACCAAACAAGAGAAGAACTCTGYTTGG | |

**B**

## Amplicon coverage of the HPIV3 genome



**Figure 1.** Primers used for amplicon generation for full genome sequencing (**a**) and the position of the amplicons along the human parainfluenza viruses type 3 (HPIV3) genome. 12 primer sets (**a**) were designed and used to generate overlapping amplicons covering the entire HPIV3 genome, as shown in (**b**).

denaturation step at 94°C, and 35 cycles of denaturation (94°C for 15s), annealing (55°C for 30s) and extension (68°C 3min 30s). After the final extension step (68°C for 5min) the reaction was held at 4°C. Following amplification the products were ran on a 1% agarose gel for confirmation and purified following the Epoch Life Science Quick Protocol for EcoSpin All-in-one Mini Spin Columns (1920-050/250 Epoch Life Science).

### Reference sequence and validation of the pipeline

Two of the isolates, MK9 and 153, were first sequenced by Sanger sequencing to validate the NGS sequencing pipeline. RNA was extracted and amplicons were generated as described above. Primers were originally designed using the genscript sequencing primer design tool. These were used for Sanger sequencing (Applied Biosystems 3730xl DNA Analyser

(Department of Biochemistry, University of Cambridge)) together with the amplification primers (See Dataset 1[20] for primer sequences, manufactured by Eurofins) aiming for overlapping amplicons of approximately 700bp each. The sequence was then aligned to a consensus sequence of the following accession numbers KF687321, KF530255, EU326526, KF530227, KF687319, KF530232, KF687317, KF530249, KF530245, KF530250, KF530229, KF530243, AB736166, KF530252, KF530236, KF530225, KF687340, KF530254, KF687318, KF530230, KF687346, KF530233, KF530242, KF530251, KF530241, KF530253, KF530257, KF530238, KF530231, KF530234, KF530247, EU424062, KF530256, FJ455842, KF687336, U51116, NC_001796.2 (see Data availability section) using Sequencher 5.4.

### NGS sequencing and analysis
The amplicons generated were combined in equimolar concentrations and sequenced on the Illumina platform (MiSeq (Clinical Translational Research Unit, Cambridge University Hospitals, NHS Foundation Trust)). Paired short reads were then processed with Trimgalore v 0.4.2 to remove the Illumina paired end library adapters as well as short and low quality reads to retain those with length >20bp and Phred scores >20. Terminal primer sequences were subsequently removed with Cutadapt v 1.14[21]. Alignment was performed with Bowtie2 v 2.2.9[22] using the Sanger sequences obtained above as reference and consensus was extracted with Samtools v 1.3.1[23]. The results were validated using Sanger sequencing of the laboratory strain MK9 and sequence 153 as well as the previously published *de novo* Ebola pipeline using QuasR v 1.20 for quality control and Spades 3.5 for assembly[24]. Variant analysis was conducted using V-Phaser2[25].

### Phylogenetics
All HPIV3 full genome sequences available on NCBI were downloaded and genomes originating from the same source and found to contain minimal variance were removed for clarity and to minimize bias. Sequences that originated from strains that were repeatedly passaged in culture or were deliberately modified, such as strain 47885, C243, 14702 and strain JS, were also removed, leaving 36 diverse full genome sequences. These, together with the 20 sequences obtained in this study were aligned in UGENE v 1.26.0 using the Muscle algorithm[26]. Subalignments were extracted using UGENE v 1.26.0. The most suitable substitution model was selected using the JModelTest 2.0 Software[27]. Position by position rates, distance matrices and maximum likelihood trees were generated using Molecular Evolutions Genetics Analysis (MEGA) software MEGA v 7[28]. Bootstrap iterations of 1000 were used for maximum likelihood tree confidence estimates. Clusters were visualized using ABGD[15]. Bayesian Markov Chain Monte Carlo (MCMC) inference model was selected using marginal likelihood estimation using path sampling (PS) and stepping stone sampling (SS) with BEAST v 1.8.4 with a chain length of one million and one hundred path steps respectively[29–32]. Tracer was used to assess convergence based on the effective sample size with 10% burn-in and effective sample size (ESS) values above 200.

Maximum clade credibility trees were generated with Tree Annotator and subsequently visualized and edited using FigTree v 1.4.3.

## Results

### Clinical samples and epidemiology of HPIV3
The number of samples that tested positive for HPIV3 on the respiratory virus panel in PHE laboratory of a major teaching hospital during the years 2011–2017 is shown in Figure 2a.

The prevalence of HPIV3 follows a cyclical pattern with peaks occurring towards the end of spring and start of summer every year. Patient demographics for each strain are summarized in Figure 2b. The patients for which strains were sequenced represented a diverse demographic with an age distribution reflecting the usual susceptibility to HPIV3 with 18/20 being below the age of five or over the age of 50. All samples were obtained from the upper respiratory tract including swabs, nasopharyngeal aspirates and tracheal aspirates. The majority of the samples (14/20) were taken from inpatients, reflecting an unavoidable sampling bias towards cases requiring admission and potential co-morbidities. Although the majority of the cases originated from one hospital (A) (12/20), the rest were from a more diverse geographical distribution reflecting the area covered by the PHE laboratory. Relevant past medical history (PMH) is shown where available (7/20) and in most cases includes patients with haematological oncology conditions such as relapsed acute lymphocytic leukaemia (ALL), immunosuppressive chemotherapy treatment (alemtuzumab) and post bone marrow transplant including allograft and volunteer unrelated donor (VUD). This reflects the immunosuppressed population where HPIV3 is known to have the highest impact[5,33]. Two chronic respiratory conditions have also been identified: cystic fibrosis (CF) and asthma. Parainfluenza viruses have known to contribute to infective exacerbations of asthma[34], particularly in paediatrics and the clinical impact of respiratory viruses on cystic fibrosis patients is well recognised[35].

### Genome coverage and variant analysis
In order to evaluate the genetic variability of UK circulating strains of HPIV3, the twenty clinical strains detailed in Figure 2a and the laboratory reference strain were sequenced by NGS on the Illumina platform. The laboratory strain and strain 153 were first sequenced by Sanger sequencing and used as reference strains for NGS pipeline validation. The depth of NGS coverage for both strains as well as the average depth achieved for all the strains sequenced is shown in Figure 3.

The depth of coverage remained consistently high apart from the 5' and 3' prime ends, confirming the robustness of the pipeline. Variant analysis was then performed using V-Phaser2 and the summary results are shown in Table 1.

### Phylogenetic analysis of the full genome sequence
In order to assess the epidemiology and evolution of HPIV3, in the context of strains circulating within the UK, a phylogenetic

## A. Total number of HPIV3 pediatric and adult cases diagnosed by PHE diagnostic laboratory between 01/11 and 08/17
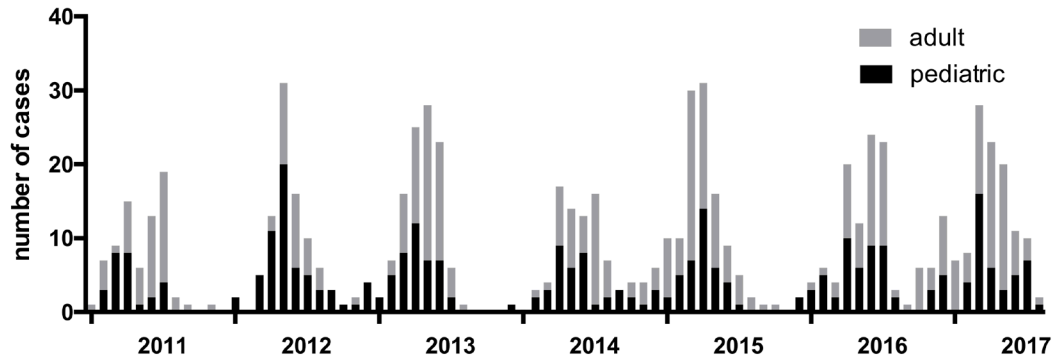


### B.

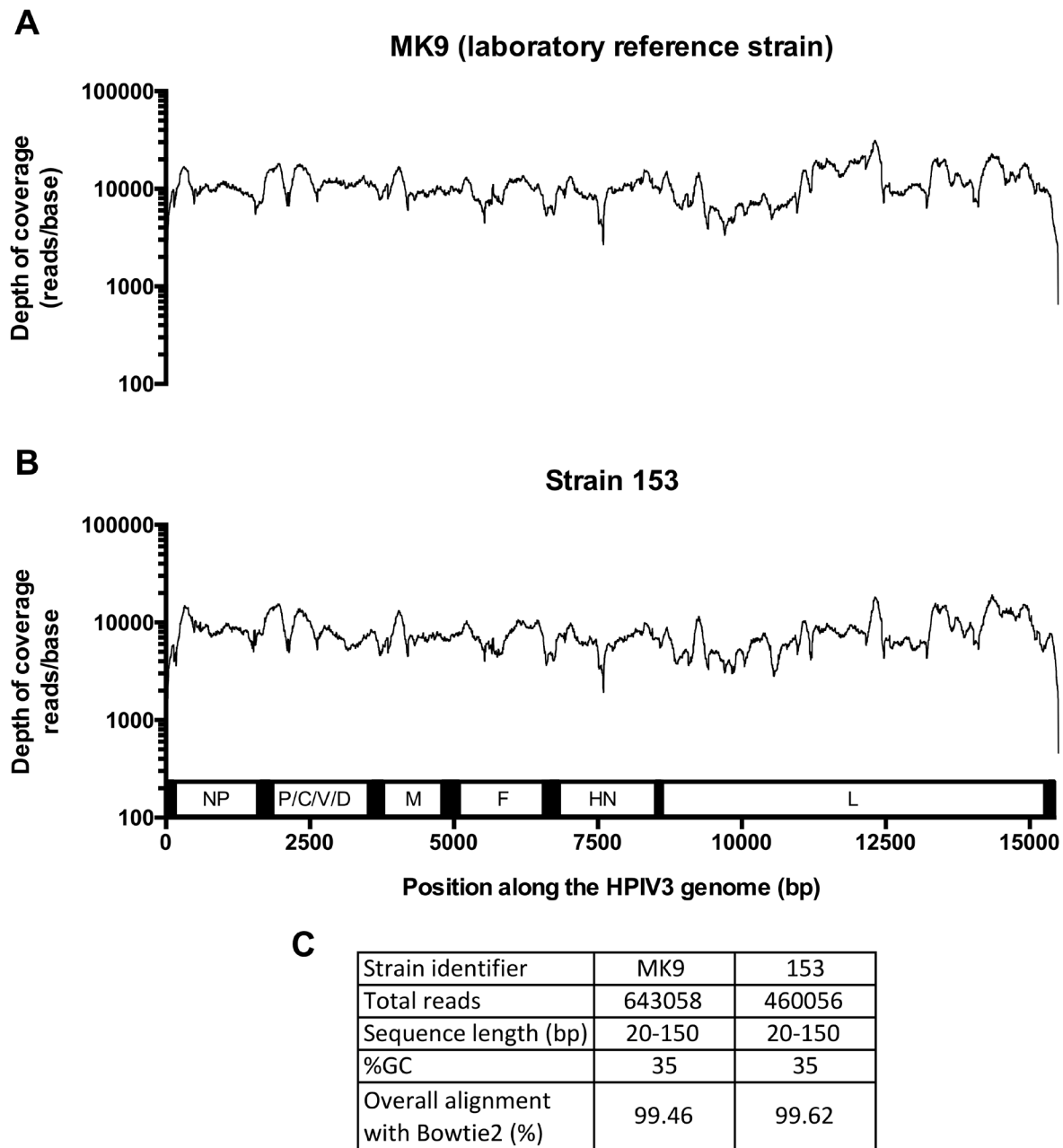| Sample | | | Patient | | | | |
|---|---|---|---|---|---|---|---|
| *Lab ID* | Date collected | Source | Sex | Age | Location | In/out patient | PMH |
| 14 | Nov-14 | URT swab | F | 1 year | hospital B | in | unknown |
| 16 | May-14 | URT swab | M | 50 years | hospital A | in | post allograft |
| 21 | Jun-14 | URT swab | F | 54 years | no data | unknown | unknown |
| 30 | Jul-14 | URT swab | M | 60 years | hospital A | out | post VUD |
| 53 | Apr-14 | URT swab | F | 58 years | hospital A | in | asthma |
| 60 | Jun-11 | URT swab | F | 56 years | hospital A | in | unknown |
| 65 | Jun-11 | URT swab | F | 54 years | hospital A | in | unknown |
| 82 | Jul-11 | URT swab | M | 2 months | hospital H | in | unknown |
| 112 | Jun-12 | URT swab | M | 84 years | hospital A | out | on alemtuzumab |
| 113 | Jul-12 | URT swab | M | 42 years | hospital A | in | unknown |
| 121 | Nov-12 | URT swab | F | 61 years | hospital A | in | post allograft |
| 122 | Dec-12 | NPA | M | 9 months | hospital HDGH | in | unknown |
| 128 | Feb-13 | NPA | F | 9 months | hospital A | in | unknown |
| 129 | Feb-13 | Tracheal aspirate | F | 4 years | hospital F | in/ICU | unknown |
| 153 | Apr-13 | NPA | F | 12 years | hospital F | in | ALL relapse |
| 180 | May-13 | NPA | F | 3 months | hospital A | out | CF |
| 362 | Mar-15 | URT swab | F | 80 years | nursing home | out | unknown |
| 371 | Apr-15 | URT swab | F | 89 years | nursing home | out | unknown |
| 390 | Feb-14 | URT swab | F | 1 year old | hospital A | in | unknown |
| 395 | Apr-15 | URT swab | M | 70 years | hospital A | in | unknown |

**Figure 2.** Total samples tested positive for human parainfluenza viruses type 3 (HPIV3) by Public Health England (PHE) diagnostic laboratory during 2011–2017 (**a**) and provenance of sequenced clinical strains (**b**). Total number of samples that have tested positive for HPIV3 in PHE diagnostic laboratory of a major teaching hospital each month are shown in (**a**) for the period 2011–2017. The data has been extracted from the local hospital database and is separated by age. The provenance of sequenced clinical strains collected between 2011 and 2015 is shown in (**b**). All samples originated from the upper airway and 12/20 samples were from hospital A. PMH = past medical history; NPA = Nasopharyngeal aspirate; URT = upper respiratory tract; VUD = volunteer unrelated donor (transplant); ALL = acute lymphocytic leukaemia.

tree of full length genome sequences available was constructed using the Maximum Likelihood method (Figure 4).

The automated barcode gap discovery analysis of the full genome sequences allowed us to define anything separated by more than a genetic distance of 0.043 as a cluster and 0.02 as a sub cluster. Therefore 2 clusters have been identified. Cluster 1 was further subdivided into subclusters, 1a and 1b, with smaller subdivisions into strains, as shown in Figure 4. It is of note that apart from strain 1b(ii) that currently only contains one full

**A**

## MK9 (laboratory reference strain)

**B**

## Strain 153



**Position along the HPIV3 genome (bp)**

**C**

| Strain identifier | MK9 | 153 |
|---|---|---|
| Total reads | 643058 | 460056 |
| Sequence length (bp) | 20-150 | 20-150 |
| %GC | 35 | 35 |
| Overall alignment with Bowtie2 (%) | 99.46 | 99.62 |

**Figure 3.** Depth of coverage achieved for laboratory strain (**A**), strain 153 (**B**) and FastQC[36] statistics for both sequences (**C**). Consistent coverage of above 1000 was achieved over the full length of the genome of both sequences excluding the very 5′ and 3′ prime ends. The length of the final sequence was 15409 base pairs, as the forward primer (26 bases) of the first amplicon, and the reverse primer of the last amplicon (27 bases) were removed in the pipeline.

genome sequence from the USA (2017) we have not observed a temporal or geographical correlation between strains. The rate of substitution/site/year has been calculated to be $4.2 \times 10^{-4}$ subs/site/year using an uncorrelated relaxed clock[37], general time reversible model (GTR) with gamma distributed rate and invariant sites[38–40] and an MCMC length of 600 million, using BEASTv1.8.4 as described in the methods. This is consistent with rates observed for other RNA viruses[41–44] The average

variability across the strains available was calculated using MEGA7 and found to be 2%.

### Analysis of variability along the genome can be used to identify a hypervariable region in HPIV3

Full genome sequencing, particularly in the context of diagnositic laboratories can be expensive and time consuming. To this end a smaller region for epidemiological and phylogentics

**Table 1. Total number of unique minor variants detected by V Phaser2 analysis.** The total number of unique minor variants detected by V Phaser 2 at each percentage level of total reads is shown. In brief, the variants are calculated both by recording the probability that a non-consensus base occurs with a greater frequency than expected by sequencing error probabilities and by analyzing he probability for non-consensus pairs of bases to co-occur given sequencing errors expected. Systematic artifacts inherent in some sequencing technologies are removed by calculating strand bias for each variant. This data is then FDR corrected and all variants with a significant (p>0.05) strand bias were excluded.

| | number of variants | | | |
|---|---|---|---|---|
| Lab ID | <1% | 1%–5% | 5%–20% | >20% |
| 14 | 7 | 0 | 2 | 0 |
| 16 | 4 | 1 | 0 | 0 |
| 21 | 7 | 3 | 1 | 1 |
| 30 | 3 | 1 | 1 | 0 |
| 53 | 12 | 7 | 0 | 0 |
| 60 | 9 | 5 | 1 | 1 |
| 65 | 9 | 5 | 1 | 0 |
| 82 | 2 | 8 | 4 | 0 |
| 112 | 0 | 6 | 13 | 0 |
| 113 | 4 | 13 | 22 | 8 |
| 121 | 0 | 3 | 2 | 0 |
| 122 | 9 | 1 | 0 | 0 |
| 128 | 1 | 1 | 4 | 1 |
| 129 | 1 | 1 | 0 | 0 |
| 153 | 4 | 3 | 0 | 0 |
| 180 | 4 | 3 | 23 | 5 |
| 362 | 7 | 0 | 1 | 0 |
| 371 | 4 | 9 | 1 | 0 |
| 390 | 2 | 1 | 0 | 0 |
| 395 | 5 | 3 | 1 | 1 |
| MK9 | 36 | 8 | 1 | 0 |

analysis was identified by calculating relative variability rates along the HPIV3 genome (Figure 5).

The site by site variability was calculated using the Tamura-Nei (TRN) model[45] of substitution with 1000 bootstrap repetitions. We have observed a peak in variability in the non-coding

region between the M gene and the F gene. For the purpose of this study this region has been defined as a region of 357 base pairs in length from position 4703 to 5160 as shown in Figure 5.

### Analysis of the hypervariable region reflects the phylogenetic profile of HPIV3

The suitability of the hypervariable region for phylogentic analysis was then evaluated by constructing a phylogentic tree and comparing it to the one obtained by using full genome sequences. The BEAST evolutionary tree for the hypervariable region can be seen in Figure 6.

The rate of substitution for this region was calculated to be $1\times10^{-3}$ subs/site/year with an average variability of 5%. This is markedly above the values calculated for the full genome sequence. Hence ABGD analysis[15] has been used to separate the sequences into subclusters corresponding to strains in the full genome analysis with a potential for finer classification for the purposes of epidemiology. The corresponding classifications are summarized in Table 2.

It is of note that only two strains (Figure 4) were not classified in the same manner by both the full genome and hypervariable analysis methods. Strain 180 has moved from subcluster 1b to subcluster 1a and strain KF530247.1 (USA2006) was noted to have moved from cluster 2 to cluster 1. It was found that strain 180 contained 3.7% ambiguous bases within its hypervariable region. This falls just below the definition of a subcluster (genetic distance of 0.04) and would potentially place this sample into two subclusters depending on the alignment. This could suggest a potential co-infection with a second strain, a feature of this sample that would not have been noted if this additional analysis were not performed. Strain KF530247.1, on the other hand contains no ambiguities and therefore the migration between clusters cannot be explained this way. It has, however been identified as a potentially recombinant sequence in a previous study[11].

Additionally, we observed that strain 1b(ii) (MF166750.1 (USA2017)) that has been identified as an emergent strain by full genome analysis (Figure 4), can now be defined as a subcluster in its own right, supporting the hypothesis that this may form a new emergent strain and subsequently subcluster of HPIV3 (Figure 6).

### Analysis of the HN coding region does not fully reflect the phylogenetic profile of HPIV3 when compared to whole genome data

In order to evaluate the current findings in the context of previously published phylogenetic classification of HPIV3[46–49] an additional analysis using only the HN coding region of the HPIV3 genome was carried out (Figure 7) and the results were compared to those obtained using whole genome sequences. Phylogenetic analysis using the F coding region was also attempted but the tree constructed did not yield clustering with high enough bootstrap values and therefore was deemed to be
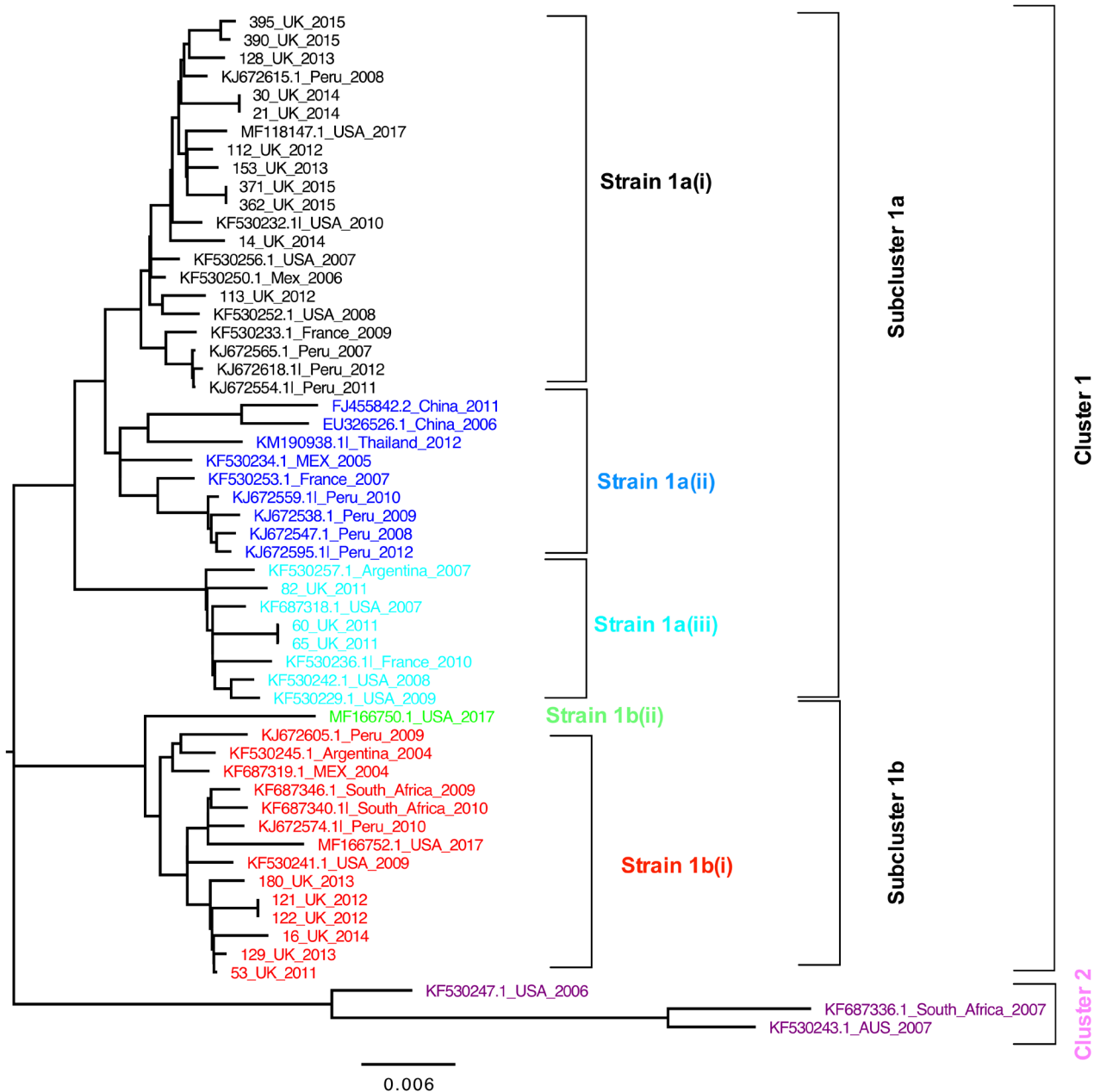
**Figure 4. Molecular Phylogenetic analysis of human parainfluenza viruses type 3 (HPIV3) full length genome by Maximum Likelihood method.** The evolutionary history was inferred by using the Maximum Likelihood method based on the General Time Reversible + I + G model and 1000 bootstrap repetitions. The tree with the highest log likelihood (-42087.04) is shown. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. The analysis involved 56 nucleotide sequences. Evolutionary analyses were conducted in MEGA7. Clusters, subclusters and strains were identified using Automatic Barcode Gap Discovery and genetic distances of 0.043 (cluster); 0.02 (subcluster) and 0.015 (strain) were identified. All strains (and cluster 2) are colored for ease of visualization and tracking.

inadequate for the current data set. It is of note that the clustering patterns observed in Figure 7 were not wholly consistent with those obtained with whole genome data (Figure 5) or the hypervariable region (Figure 6). Strain 1b(ii) was found to contain 4 sequences: KF687319.1; KF530245.1; KJ67605 and MF166750.1. When using whole genome data, only the latter was classified within this strain, whereas the other sequences were clustered within strain 1b (i).

## Discussion

In this study we have presented a robust amplicon based sequencing pipeline for the evaluation of genetic variability of circulation UK strains between the years 2011–2015. The consensus sequences generated by the pipeline were validated using both Sanger sequencing and a published *de novo* pipeline[24]. The depth of coverage obtained across all genome sequenced was consistently high (Figure 3). Variant calling analysis was
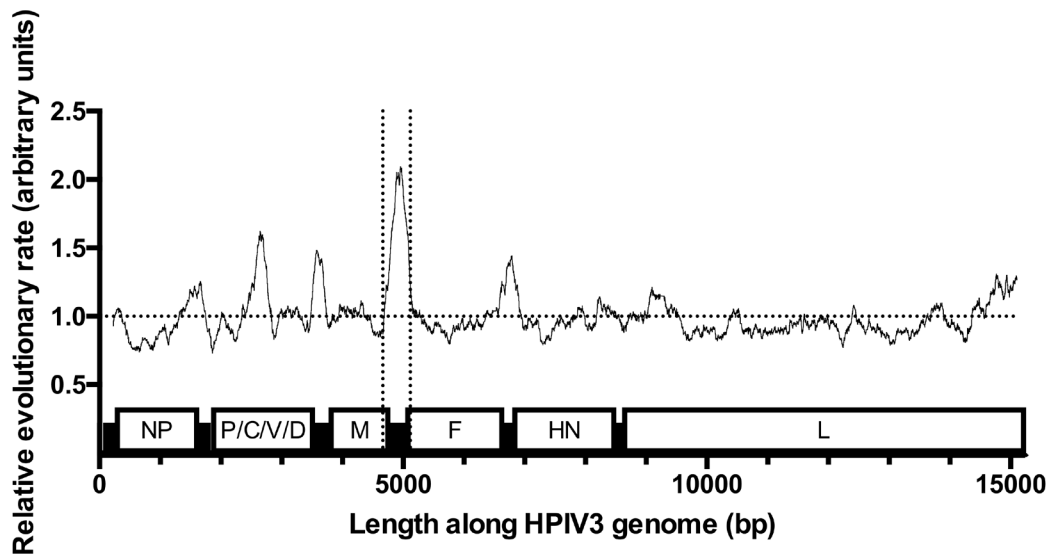
**Figure 5. Relative site by site evolutionary rate of the human parainfluenza viruses type 3 (HPIV3) genome.** Mean (relative) evolutionary rate are shown for each site next to the site number with a window of 200. These rates are scaled such that the average evolutonary rate across all sites is 1. This means that sites showing a rate < 1 are evolving slower than average, and those with a rate > 1 are evolving faster than average. These relative rate were estimated under the General Time Reversible model (+G+I). The analysis involved 56 nucleotide sequences. The position along the HPIV3 genome is shown on the x axis with the hypervariable region identified between positions 4703 to 5160. Evolutionary analyses were conducted in MEGA7.

performed using VPhaser2 and the numbers of unique variants determined did not demonstrate a high degree of within host variability. This may be a reflection of the amplicon based amplification method used for this pipeline as capture-based methods have been shown to provide a more robust minor variant amplification approach[50,51]. However amplicon based methods have previously been used successfully for variant calling in a clinical setting[52] and have been considered to be effective for minority variable identification. On the other hand, as all the sequenced strains were grown from clinical samples in immortalized cell tissue culture, it most likely reflects the genetic bottleneck due to this method[53].

The clinical strains sequenced were drawn from a wide demographic with an inevitable bias towards inpatient admissions due to increased viral testing in these circumstances. The sample demographic reflected known populations with a known severe impact by HPIV3. 18 out of 20 strains were obtained from patients aged below 5 or over 50. Although clinical data was only available in seven cases, two patients were identified as having chronic respiratory conditions and five were immuno-suppressed due to a haemato-oncological condition. Although this sample formed a good representation of the population most commonly severely affected by HPIV3[33–35], further studies focusing on the occurrence of HPIV3 in the otherwise healthy cohort could enhance the knowledge of its variability.

To this end the phylogenetic analysis of HPIV3 UK sequences obtained was conducted in the context of other full genome sequences currently available in the NCBI database. We have excluded all duplicate and highly similar sequences as well as

significantly culture adapted strains. Consequently we have observed only two clusters, which contrasts with three previously identified in literature[54]. This is most likely due to the exclusion of strains from the last century that are known to be heavily culture adapted and have therefore not been included in this study. We have observed that cluster 1 (Figure 4) contains most strains that are currently circulating including all the UK strains sequenced in this study. Cluster 1 can further be subdivided into two subclusters with UK strains falling into both rather than forming a UK specific subcluster or strain.

We have not noted any significant geographical or temporal correlation between strains and clusters using the limited number of diverse sequences available to us (Figure 4). The majority of the strains were observed at different geographical locations and with a wide temporal spread. This reflects that HPIV3 is a global problem that has been identified as an important respiratory pathogen in all countries conducting respiratory virus surveillance[4]. The rate of substitution for HPIV3 ($4.2 \times 10^{-4}$ subs/site/year) is consistent with other RNA viruses[41,42] and would predict a development of a new subcluster roughly every 50 years and a new cluster every 107 years. We have observed a potential new emerging strain MF166750.1 (USA2017), and a strain (1a(ii)) that has not appeared since 2012 (Figure 4). Although the rate of evolution, in itself, would be insufficient to account for the yearly rise in cases seen, it does reflect a number of globally circulating strains as well as regularly emerging new ones.

Limited geographical clustering observed in this analysis was most likely due to the scarcity of diverse sequences from
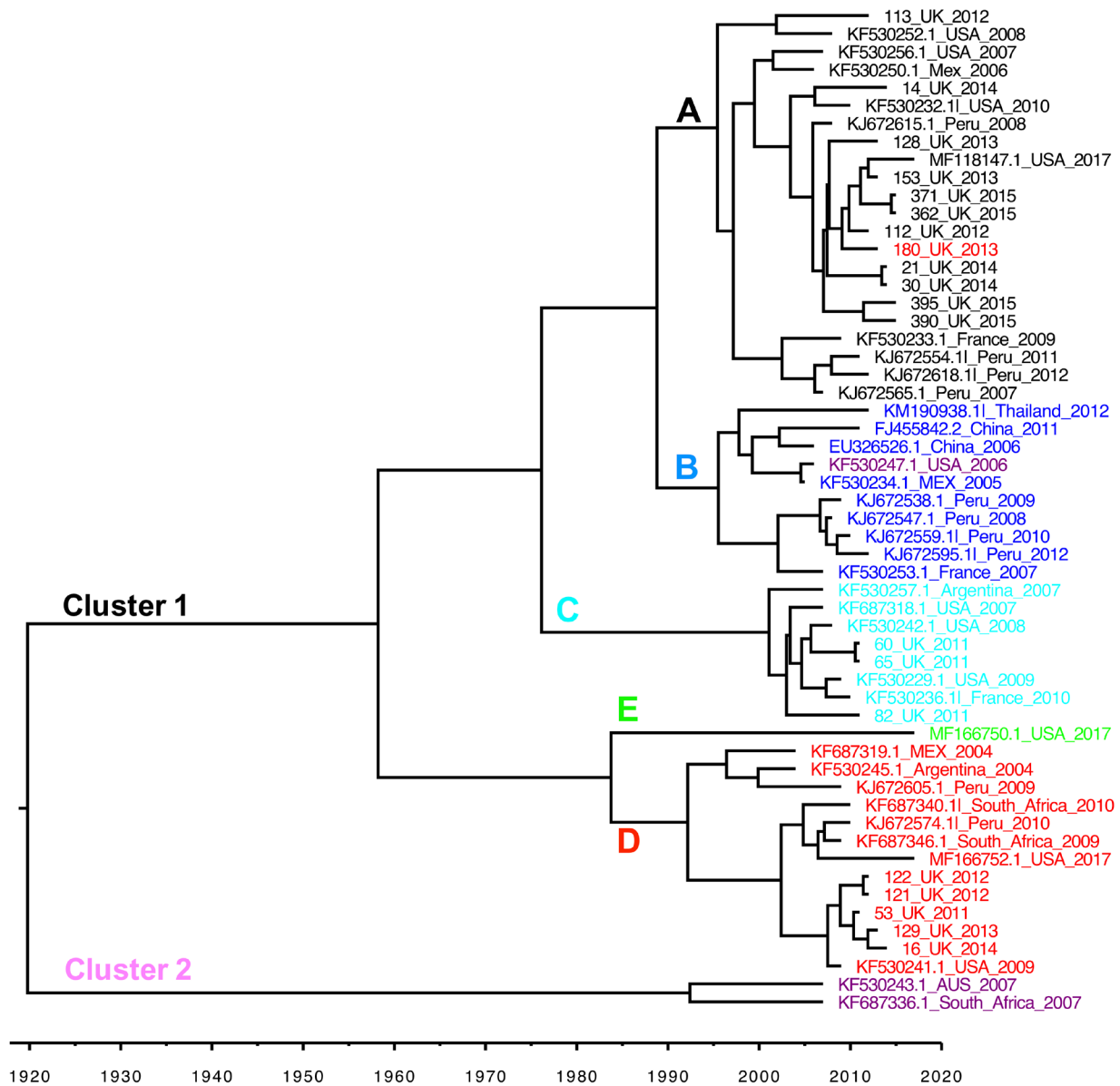
**Figure 6. Molecular Phylogenetic analysis of human parainfluenza viruses type 3 (HPIV3) hypervariable region by Bayesian Phylogenetics using BEAST.** The evolutionary history was inferred by using Bayesian Phylogenetics based on the TRN +I model using BEAST v1.8.4. The tree with the highest log likelihood (-2020.47) using path sampling and stepping stone analysis is shown. A strict clock and a constant coalescent prior were used. The MCMC length was 10,000,000. Convergence was assessed with Tracer and the maximum clade credibility tree was generated with Tree Annotator. Dates of strain emergence (in years) are shown in the figure legend. Automatic Barcode Gap Discovery was used to analyse genetic distances and 0.1 (cluster) and 0.04 (subcluster) were defined. Subclusters A-E are shown next to their respective branches. All strains and cluster and subcluster labels are colored identically to the phylogenetic analysis using full genome (Figure 4) to demonstrate near identical clustering patterns.

specific geographical regions, particularly for South/East Asia. Over the last few years, this number has increased, partially due to decreased cost and greater access to next generation sequencing[55] and a greater interest in viral metagenomics of respiratory infections[56]. However the bulk of whole genome sequences available centers on a small number of episodes or studies, yielding clusters of highly similar or identical genomes[57]. Most historical sequences used for previous phylogenetic analyses[46–48], omitted in this study, have been propagated and maintained in cell culture prior to sequencing[58–60] Cell culture adaptation of HPIV3 has been well studied both phenotypically, where culture adapted strains were shown to produce larger plaques in cell culture as well as genotypically, with a particular focus on HN and F coding regions, where it has been shown that

**Table 2. Subdivisions identified by Automatic barcode gap discovery (ABGD) analysis of whole genome sequences of HPIV3 and hypervariable region of human parainfluenza viruses type 3 (HPIV3).** The corresponding subdivisions into clusters, subclusters and strains, identified by ABGD for whole genome sequences (Figure 4 and the hypervariable region of HPIV3 (Figure 6) are summarized.

| Subdivisions identified by ABGD analysis of | | | | |
|---|---|---|---|---|
| Whole genome | | | Hypervariable region | |
| Cluster | Subcluster | Strain | Cluster | Subcluster |
| 1 | 1a | 1a(i) | 1 | A |
| | | 1a(ii) | | B |
| | | 1a(iii) | | C |
| | 1b | 1b(i) | | D |
| | | 1b(ii) | | E |
| 2 | | | 2 | |

different adaptations are required for growth in cell culture as opposed to the natural host[61–63].

It could be argued that some degree of adaptation was also inherent in the strains used in this study, as they have undergone one passage in PLC/PRF/5 cells prior to sequencing. However the sequences obtained were minimally adapted compared to historical strains maintained in cell culture over many passages[64]. Additionally these clinical sequences were found to be phylogenetically diverse and to cluster with other clinical strains of HPIV3, whereas significantly culture adapted strains have previously been shown to form a separate cluster, as described above. An alternative approach to whole genome sequencing, using the small hypervariable region, easily sequenced from clinical samples was also explored in this paper, and could be used for phylogenetic typing of HPIV3 where whole genome data cannot be obtained.

We have demonstrated that a salient phylogenetic and epidemiological analysis can be performed on a short hypervariable region of the HPIV3 genome described in this study. Although an overlapping region has been used previously for outbreak monitoring, it was a shorter segment of 244 bases, position 4880 to 5124[8,9]. This would place the primer sites within the hypervariable region itself (defined here as position 4703 to 5160, Figure 3), potentially limiting amplification success rate. In the current analysis the region identified is flanked by two regions of reduced variability, creating ideal locations for primer localization (Figure 5).

A faster substitution rate of the hypervariable region ($1 \times 10^{-3}$) has facilitated the division into 2 clusters and 5 further subclusters. All strains apart from two, have been shown to cluster

in an identical pattern to that identified in the full genome analysis. MF166750.1 (USA2017) has been identified as forming its own subcluster, which was consistent with full genome analysis, where it was identified as an emerging strain (Figure 2). This is particularly salient in a clinical environment where whole genome sequencing may not be feasible due to the quality and storage of the clinical samples. The costs and turn around times associated with next generation sequencing are often prohibitive in a diagnostic laboratory. The comparatively short hypervariable fragment can be sequenced by Sanger sequencing, decreasing costs, turn around times and removing the need for complex bioinformatics analysis as demonstrated in the companion paper[65].

It is of note that only three full genome strains were found to fall into cluster 2 when full genome phylogenetic analysis was conducted and one of the strains (KF530247.1) moved to cluster 1 when only the hypervariable region was used for analysis (Figure 2 and Figure 4). KF530247.1 was previously identified in literature as a potential recombinant strain, where phylogenetic analysis using its HN genome did not correlate with that of the F genome[11]. It is important to note that this strain and the two remaining strains form part of the same sequencing project, accession number PRJNA73055, however other strains from the same project behaved as predicted in our analysis. Hence, this likely reflects the scarcity of full genome data available for the analysis of HPIV3. Overall, the geographical distribution of sequences available was noted to be heavily biased towards the northern hemisphere and the Americas. This highlights the necessity to conduct further epidemiological surveillance of HPIV3 as well as the utility of a smaller part of the genome that could be used for this purpose, as identified in this study.

The second sample that did not cluster in the same fashion as that predicted by full genome analysis was UK strain 180. Further analysis of the sequence revealed that there were 3.7% ambiguous bases in the hypervariable region of this strain. This ambiguity could influence the subcluster (genetic distance 0.043) in which this strain would have been grouped. As a number of different strains have been shown to circulate globally each year, we could hypothesise that this is a case of dual infection that was identified when the analyses of the hypervariable region and the full genome were compared.

As the HN coding region has previously been used for phylogenetic characterization of HPIV3[46–49] this approach was also applied to the current data set (Figure 7). It is interesting to note that the resulting clustering patterns were not wholly consistent with those obtained by whole genome sequencing. Three strains (KF687319.1; KF530245.1; KJ67605) were classified differently (Figure 7). However both strains 180 and strain KF530271.1 were classified in agreement with whole genome analysis (Figure 5). The discrepant strains in this case were sourced from different projects, have not been previously identified in literature as atypically clustering, and did not contain sufficient ambiguous bases that could have potentially influenced the classification, as described above for strains 180 and
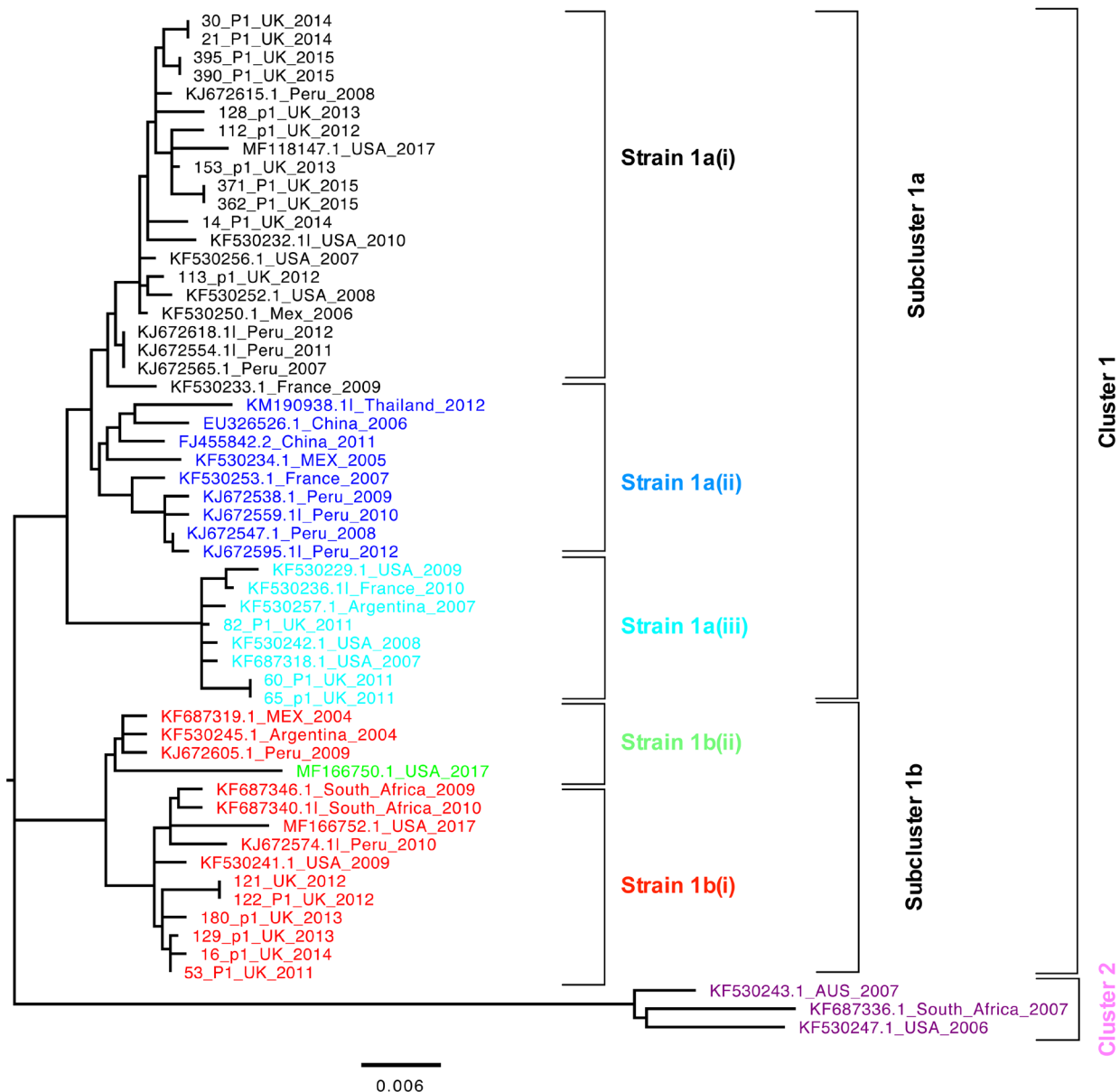
**Figure 7. Molecular Phylogenetic analysis of human parainfluenza viruses type 3 (HPIV3) HN coding region by Maximum Likelihood method.** The evolutionary history was inferred by using the Maximum Likelihood method based on the TRN +G model and 1000 bootstrap repetitions. The tree with the highest log likelihood (-4490.43) is shown. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. The analysis involved 56 nucleotide sequences. Evolutionary analyses were conducted in MEGA7. Clusters, subclusters and strains were identified using Automatic Barcode Gap Discovery and genetic distances of 0.043 (cluster); 0.02 (subcluster) and 0.01 (strain) were identified. All strains (and cluster 2) are colored consistent with Figure 4 for ease of visualization and tracking.

KF530271.1 Unfortunately, without further understanding of the connection between genotype and clinical and evolutionary phenotype of HPIV3, it is impossible to infer which analysis produced more biologically salient results. However we have shown that phylogenetic analysis using HN did not fully reflect the clustering obtained using whole genome data for this data set and this was matched more closely by the hypervariable region identified in this study.

In conclusion, in this study we have presented a first genetic analysis of whole genome sequences of circulating UK strains of HPIV3 in the period of 2011–2015. We have concluded that, consistent with previous data, the circulating strains fall into the same cluster but do not form a particular subcluster or strain[11–13,54]. Within the constraints of the amount of sequences available we have not observed a temporal or geographical correlation with a particular strain of HPIV3. However the number of full genome

HPIV3 sequences is limited, necessitating further epidemiological investigation both in a laboratory and clinical setting. To this end, we have identified a small hypervariable region in the HPIV3 genome and have shown that, in the absence of full genome data, this region can be used for epidemiological surveillance of HPIV3.

## Data availability

Raw data and details of primers used for Sanger sequencing available on Open Science Framework. Dataset 1: UK circulating strains of human parainfluenza 3: an amplicon based next generation sequencing method and phylogenetic analysis. https://doi.org/10.17605/OSF.IO/SUQ95[20]

All HPIV3 sequences have been uploaded to NCBI with the following accession numbers:

HPIV3/UK/ **65**/06/2011 MH678674

HPIV3/UK/ **112**/06/2012 MH678675

HPIV3/UK/ **113**/07/2012 MH678676

HPIV3/UK/ **121**/11/2012 MH678677

HPIV3/UK/ **128**/02/2013 MH678678

HPIV3/UK/ **129**/02/2013 MH678679

HPIV3/UK/ **153**/04/2013 MH678680

HPIV3/UK/ **180**/05/2013 MH678681

HPIV3/UK/ref/ **MK9** MH678682

HPIV3/UK/ **14**/11/2014 MH678683

HPIV3/UK/ **16**/05/2014 MH678684

HPIV3/UK/ **21**/06/2014 MH678685

HPIV3/UK/ **30**/07/2014 MH678686

HPIV3/UK/ **60**/06/2011 MH678687

HPIV3/UK/ **82**/07/2011 MH678688

HPIV3/UK/ **122**/12/2012 MH678689

HPIV3/UK/ **362**/03/2015 MH678690

HPIV3/UK/ **371**/04/2015 MH678691

HPIV3/UK/ **390**/02/2014 MH678692

HPIV3/UK/ **395**/04/2015 MH678693

HPIV3/UK/ **53**/04/2014 MH678694

All uploaded sequences can be accessed together via PopSet: 1470015825

## References

1.  Henrickson KJ: **Parainfluenza viruses.** *Clin Microbiol Rev.* 2003; **16**(2): 242–64.
    **PubMed Abstract** | **Publisher Full Text** | **Free Full Text**

2.  Vainionpää R, Hyypiä T: **Biology of parainfluenza viruses.** *Clin Microbiol Rev.* 1994; **7**(2): 265–75.
    **PubMed Abstract** | **Publisher Full Text** | **Free Full Text**

3.  Zhao H, Harris RJ, Ellis J, *et al.*: **Epidemiology of parainfluenza infection in England and Wales, 1998-2013: any evidence of change?** *Epidemiol Infect.* 2017; **145**(6): 1210–1220.
    **PubMed Abstract** | **Publisher Full Text**

4.  Weinberg GA, Hall CB, Iwane MK, *et al.*: **Parainfluenza virus infection of young children: estimates of the population-based burden of hospitalization.** *J Pediatr.* 2009; **154**(5): 694–9.
    **PubMed Abstract** | **Publisher Full Text**

5.  Shah DP, Shah PK, Azzi JM, *et al.*: **Parainfluenza virus infections in hematopoietic cell transplant recipients and hematologic malignancy patients: A systematic review.** *Cancer Lett.* 2016; **370**(2): 358–364.
    **PubMed Abstract** | **Publisher Full Text** | **Free Full Text**

6.  Seo S, Xie H, Campbell AP, *et al.*: **Parainfluenza virus lower respiratory tract disease after hematopoietic cell transplant: viral detection in the lung predicts outcome.** *Clin Infect Dis.* 2014; **58**(10): 1357–1368.
    **PubMed Abstract** | **Publisher Full Text** | **Free Full Text**

7.  Brady MT, Evans J, Cuartas J: **Survival and disinfection of parainfluenza viruses on environmental surfaces.** *Am J Infect Control.* 1990; **18**(1): 18–23.
    **PubMed Abstract** | **Publisher Full Text**

8.  Jalal H, Bibby DF, Bennett J, *et al.*: **Molecular investigations of an outbreak of parainfluenza virus type 3 and respiratory syncytial virus infections in a hematology unit.** *J Clin Microbiol.* 2007; **45**(6): 1690–6.
    **PubMed Abstract** | **Publisher Full Text** | **Free Full Text**

9.  Zambon M, Bull T, Sadler CJ, *et al.*: **Molecular epidemiology of two consecutive outbreaks of parainfluenza 3 in a bone marrow transplant unit.** *J Clin Microbiol.* 1998; **36**(8): 2289–2293.
    **PubMed Abstract** | **Free Full Text**

10. Almajhdi FN, Alshaman MS, Amer HM: **Molecular characterization and phylogenetic analysis of human parainfluenza virus type 3 isolated from Saudi Arabia.** *J Med Virol.* 2012; **84**(8): 1304–11.
    **PubMed Abstract** | **Publisher Full Text**

11. Košutić-Gulija T, Slovic A, Ljubin-Sternak S, *et al.*: **Genetic analysis of human parainfluenza virus type 3 obtained in Croatia, 2011–2015.** *J Med Microbiol.* 2017; **66**(4): 502–510.
    **PubMed Abstract** | **Publisher Full Text**

12. Mao N, Ji Y, Xie Z, *et al.*: **Human parainfluenza virus-associated respiratory tract infection among children and genetic analysis of HPIV-3 strains in Beijing, China.** *PLoS One.* 2012; **7**(8): e43893.
    **PubMed Abstract** | **Publisher Full Text** | **Free Full Text**

13. Goya S, Mistchenko AS, Viegas M: **Phylogenetic and molecular analyses of human parainfluenza type 3 virus in Buenos Aires, Argentina, between 2009 and 2013: The emergence of new genetic lineages.** *Infect Genet Evol.* 2016; **39**: 85–91.
    **PubMed Abstract** | **Publisher Full Text**

14. Piralla A, Percivalle E, Di Cesare-Merlone A, *et al.*: **Multicluster nosocomial outbreak of parainfluenza virus type 3 infection in a pediatric oncohematology unit: a phylogenetic study.** *Haematologica.* 2009; **94**(6): 833–9.
    **PubMed Abstract** | **Publisher Full Text** | **Free Full Text**

15. Puillandre N, Lambert A, Brouillet S, *et al.*: **ABGD, Automatic Barcode Gap Discovery for primary species delimitation.** *Mol Ecol.* 2012; **21**(8): 1864–1877.
    **PubMed Abstract** | **Publisher Full Text**

16. Lingemann M, Surman S, Amaro-Carambot E, *et al.*: **The aberrant gene-end transcription signal of the matrix M gene of human parainfluenza virus type 3 downregulates fusion F protein expression and the F-specific antibody response *in vivo*.** *J Virol.* 2015; **89**(6): 3318–31.
**PubMed Abstract** | **Publisher Full Text** | **Free Full Text**

17. Kato A, Kiyotani K, Hasan MK, *et al.*: **Sendai virus gene start signals are not equivalent in reinitiation capacity: moderation at the fusion protein gene.** *J Virol.* 1999; **73**(11): 9237–46.
**PubMed Abstract** | **Free Full Text**

18. Greninger AL, Zerr DM, Qin X, *et al.*: **Rapid Metagenomic Next-Generation Sequencing during an Investigation of Hospital-Acquired Human Parainfluenza Virus 3 Infections.** *J Clin Microbiol.* 2016; **55**(1): 177–182.
**PubMed Abstract** | **Publisher Full Text** | **Free Full Text**

19. Smielewska A, Emmott E, Goodfellow I, *et al.*: **In vitro sensitivity of human parainfluenza 3 clinical isolates to ribavirin, favipiravir and zanamivir.** *J Clin Virol.* 2018; **102**: 19–26.
**PubMed Abstract** | **Publisher Full Text**

20. Smielewska A: **UK Circulating Strains of Human Parainfluenza 3: An Amplicon Based next Generation Sequencing Method and Phylogenetic Analysis.** *Open Science Framework.* 2018.
**http://www.doi.org/10.17605/OSF.IO/SUQ95**

21. Martin M: **Cutadapt removes adapter sequences from high-throughput sequencing reads.** *EMBnet J.* 2011; **17**(1): 10.
**Publisher Full Text**

22. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2.** *Nat Methods.* 2012; **9**(4): 357–359.
**PubMed Abstract** | **Publisher Full Text** | **Free Full Text**

23. Li H, Handsaker B, Wysoker A, *et al.*: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics.* 2009; **25**(16): 2078–2079.
**PubMed Abstract** | **Publisher Full Text** | **Free Full Text**

24. Arias A, Watson SJ, Asogun D, *et al.*: **Rapid outbreak sequencing of Ebola virus in Sierra Leone identifies transmission chains linked to sporadic cases.** *Virus Evol.* 2016; **2**(1): vew016.
**PubMed Abstract** | **Publisher Full Text** | **Free Full Text**

25. Yang X, Charlebois P, Macalalad A, *et al.*: **V-Phaser 2: variant inference for viral populations.** *BMC Genomics.* 2013; **14**: 674.
**PubMed Abstract** | **Publisher Full Text** | **Free Full Text**

26. Okonechnikov K, Golosova O, Fursov M: **Unipro UGENE: a unified bioinformatics toolkit.** *Bioinformatics.* 2012; **28**(8): 1166–1167.
**PubMed Abstract** | **Publisher Full Text**

27. Darriba D, Taboada GL, Doallo R, *et al.*: **jModelTest 2: more models, new heuristics and parallel computing.** *Nat Methods.* 2012; **9**(8): 772.
**PubMed Abstract** | **Publisher Full Text** | **Free Full Text**

28. Tamura K, Dudley J, Nei M, *et al.*: **MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0.** *Mol Biol Evol.* 2007; **24**(8): 1596–1599.
**PubMed Abstract** | **Publisher Full Text**

29. Bryant JE, Holmes EC, Barrett AD: **Out of Africa: a molecular perspective on the introduction of yellow fever virus into the Americas.** *PLoS Pathog.* 2007; **3**(5): e75.
**PubMed Abstract** | **Publisher Full Text** | **Free Full Text**

30. Xie W, Lewis PO, Fan Y, *et al.*: **Improving marginal likelihood estimation for Bayesian phylogenetic model selection.** *Syst Biol.* 2011; **60**(2): 150–60.
**PubMed Abstract** | **Publisher Full Text** | **Free Full Text**

31. Drummond AJ, Suchard MA, Xie D, *et al.*: **Bayesian phylogenetics with BEAUti and the BEAST 1.7.** *Mol Biol Evol.* 2012; **29**(8): 1969–73.
**PubMed Abstract** | **Publisher Full Text** | **Free Full Text**

32. Ayres DL, Darling A, Zwickl DJ, *et al.*: **BEAGLE: an application programming interface and high-performance computing library for statistical phylogenetics.** *Syst Biol.* 2012; **61**(1): 170–173.
**PubMed Abstract** | **Publisher Full Text** | **Free Full Text**

33. Nichols WG, Corey L, Gooley T: **Parainfluenza virus infections after hematopoietic stem cell transplantation: risk factors, response to antiviral therapy, and effect on transplant outcome.** *Blood.* 2001; **98**(3): 573–578.
**PubMed Abstract** | **Publisher Full Text**

34. Lewandowska-Polak A, Brauncajs M, Paradowska E, *et al.*: **Human parainfluenza virus type 3 (HPIV3) induces production of IFNγ and RANTES in human nasal epithelial cells (HNECs).** *J Inflamm (Lond).* 2015; **12**: 16.
**PubMed Abstract** | **Publisher Full Text** | **Free Full Text**

35. Flight W, Jones A: **The diagnosis and management of respiratory viral infections in cystic fibrosis.** *Expert Rev Respir Med.* 2017; **11**(3): 221–227.
**PubMed Abstract** | **Publisher Full Text**

36. Drummond AJ, Ho SY, Phillips MJ, *et al.*: **Relaxed phylogenetics and dating with confidence.** *PLoS Biol.* 2006; **4**(5): e88.
**PubMed Abstract** | **Publisher Full Text** | **Free Full Text**

37. Andrews S: **FastQC A Quality Control tool for High Throughput Sequence Data.** 2018.
**Reference Source**

38. Tavaré S: **Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences.** In *American Mathematical Society: Lectures on Mathematics in the Life Sciences.* Amer Mathematical Society. 1986; **17**: 57–86.
**Reference Source**

39. Yang Z: **Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods.** *J Mol Evol.* 1994; **39**(3): 306–14.
**PubMed Abstract** | **Publisher Full Text**

40. Ferreira MAR, Suchard MA: **Bayesian analysis of elapsed times in continuous-time Markov chains.** *Can J Stat.* 2008; **36**(3): 355–368.
**Publisher Full Text**

41. Jenkins GM, Rambaut A, Pybus OG, *et al.*: **Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis.** *J Mol Evol.* 2002; **54**(2): 156–165.
**PubMed Abstract** | **Publisher Full Text**

42. Sanjuán R, Nebot MR, Chirico N, *et al.*: **Viral mutation rates.** *J Virol.* 2010; **84**(19): 9733–48.
**PubMed Abstract** | **Publisher Full Text** | **Free Full Text**

43. Figlerowicz M, Alejska M, Kurzyńska-Kokorniak A, *et al.*: **Genetic variability: the key problem in the prevention and therapy of RNA-based virus infections.** *Med Res Rev.* 2003; **23**(4): 488–518.
**PubMed Abstract** | **Publisher Full Text**

44. Tan L, Coenjaerts FE, Houspie L, *et al.*: **The comparative genomics of human respiratory syncytial virus subgroups A and B: genetic variability and molecular evolutionary dynamics.** *J Virol.* 2013; **87**(14): 8213–26.
**PubMed Abstract** | **Publisher Full Text** | **Free Full Text**

45. Tamura K, Nei M: **Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees.** *Mol Biol Evol.* 1993; **10**(3): 512–526.
**PubMed Abstract** | **Publisher Full Text**

46. Almajhdi FN: **Hemagglutinin-neuraminidase gene sequence-based reclassification of human parainfluenza virus 3 variants.** *Intervirology.* 2015; **58**(1): 35–40.
**PubMed Abstract** | **Publisher Full Text**

47. Mao N, Ji Y, Xie Z, *et al.*: **Human parainfluenza virus-associated respiratory tract infection among children and genetic analysis of HPIV-3 strains in Beijing, China.** *PLoS One.* 2012; **7**(8): e43893.
**PubMed Abstract** | **Publisher Full Text** | **Free Full Text**

48. Košutić-Gulija T, Slovic A, Ljubin-Sternak S, *et al.*: **Genetic analysis of human parainfluenza virus type 3 obtained in Croatia, 2011–2015.** *J Med Microbiol.* 2017; **66**(4): 502–510.
**PubMed Abstract** | **Publisher Full Text**

49. Goya S, Mistchenko AS, Viegas M: **Phylogenetic and molecular analyses of human parainfluenza type 3 virus in Buenos Aires, Argentina, between 2009 and 2013: The emergence of new genetic lineages.** *Infect Genet Evol.* 2016; **39**: 85–91.
**PubMed Abstract** | **Publisher Full Text**

50. García-García G, Baux D, Faugère V, *et al.*: **Assessment of the latest NGS enrichment capture methods in clinical context.** *Sci Rep.* 2016; **6**: 20948.
**PubMed Abstract** | **Publisher Full Text** | **Free Full Text**

51. Samorodnitsky E, Jewell BM, Hagopian R, *et al.*: **Evaluation of Hybridization Capture Versus Amplicon-Based Methods for Whole-Exome Sequencing.** *Hum Mutat.* 2015; **36**(9): 903–914.
**PubMed Abstract** | **Publisher Full Text** | **Free Full Text**

52. Huber M, Metzner KJ, Geissberger FD, *et al.*: **MinVar: A rapid and versatile tool for HIV-1 drug resistance genotyping by deep sequencing.** *J Virol Methods.* 2017; **240**: 7–13.
**PubMed Abstract** | **Publisher Full Text**

53. Schaap-Nutt A, Liesman R, Bartlett EJ, *et al.*: **Human parainfluenza virus serotypes differ in their kinetics of replication and cytokine secretion in human tracheobronchial airway epithelium.** *Virology.* 2012; **433**(2): 320–8.
**PubMed Abstract** | **Publisher Full Text** | **Free Full Text**

54. Almajhdi FN: **Hemagglutinin-neuraminidase gene sequence-based reclassification of human parainfluenza virus 3 variants.** *Intervirology.* 2015; **58**(1): 35–40.
**PubMed Abstract** | **Publisher Full Text**

55. Kothari A, Burgess MJ, Crescencio JCR, *et al.*: **The role of next generation sequencing in infection prevention in human parainfluenza virus 3 infections in immunocompromised patients.** *J Clin Virol.* 2017; **92**: 53–55.
**PubMed Abstract** | **Publisher Full Text** | **Free Full Text**

56. Madi N, Al-Nakib W, Mustafa AS, *et al.*: **Metagenomic analysis of viral diversity in respiratory samples from patients with respiratory tract infections in Kuwait.** *J Med Virol.* 2018; **90**(3): 412–420.
**PubMed Abstract** | **Publisher Full Text**

57. Greninger AL, Zerr DM, Qin X, *et al.*: **Rapid Metagenomic Next-Generation Sequencing during an Investigation of Hospital-Acquired Human Parainfluenza Virus 3 Infections.** *J Clin Microbiol.* 2016; **55**(1): 177–182.
**PubMed Abstract** | **Publisher Full Text** | **Free Full Text**

58. Galinski MS, Mink MA, Pons MW: **Molecular cloning and sequence analysis of the human parainfluenza 3 virus gene encoding the L protein.** *Virology.* 1988; **165**(2): 499–510.
**PubMed Abstract** | **Publisher Full Text**

59. Galinski MS, Wechsler SL: **The Molecular Biology of the *Paramyxovirus* Genus.** In *The Paramyxoviruses.* (ed. Kingsbury, D. W.) Springer US, 1991; 41–82.
**Publisher Full Text**

60. Belshe RB, Hissom FK: **Cold adaptation of parainfluenza virus type 3: induction of three phenotypic markers.** *J Med Virol.* 1982; **10**(4): 235–242.
**PubMed Abstract** | **Publisher Full Text**

61. Moscona A: **Entry of parainfluenza virus into cells as a target for interrupting childhood respiratory disease.** *J Clin Invest.* 2005; **115**(7): 1688–1698 .
**PubMed Abstract** | **Publisher Full Text** | **Free Full Text**

62. Palmer SG, DeVito I, Jenkins SG, *et al.*: **Circulating clinical strains of human parainfluenza virus reveal viral entry requirements for *in vivo* infection.** *J Virol.* 2014; **88**(22): 13495–502.
**PubMed Abstract** | **Publisher Full Text** | **Free Full Text**

63. Palmer SG, Porotto M, Palermo LM, *et al.*: **Adaptation of human parainfluenza virus to airway epithelium reveals fusion properties required for growth in host tissue.** *mBio.* 2012; **3**(3): pii: e00137-12.
**PubMed Abstract** | **Publisher Full Text** | **Free Full Text**

64. Smielewska A, Emmott E, Goodfellow I, *et al.*: ***In vitro* sensitivity of human parainfluenza 3 clinical isolates to ribavirin, favipiravir and zanamivir.** *J Clin Virol.* 2018; **102**: 19–26.
**PubMed Abstract** | **Publisher Full Text**

65. Smielewska A, Pearson C, Popay A, *et al.*: **Unrecognised Outbreak: Human Parainfluenza Virus Infections in a pediatric oncology unit. A new diagnostic PCR and virus monitoring system may allow early detection of future outbreaks.** *Wellcome Open Res.* 2018.
**Publisher Full Text**

# Open Peer Review

## Current Referee Status: ✔ ✔

---

**Version 2**

Referee Report 03 December 2018

https://doi.org/10.21956/wellcomeopenres.16293.r34334

✔ **Hirokazu Kimura** [1,2]

[1] School of Medical Technology, Faculty of Health Sciences, Gunma Paz University, Gunma, Japan
[2] Graduate School of Health Science, Gunma Paz University, Gunma, Japan

The authors addressed all my comments. Thus, the revised manuscript is acceptable for indexing.

*Competing Interests:* No competing interests were disclosed.

**I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

Referee Report 28 November 2018

https://doi.org/10.21956/wellcomeopenres.16293.r34333

✔ **Julian Wei-Tze Tang** [1,2]

[1] Clinical Microbiology, University Hospitals of Leicester NHS Trust, Leicester, UK
[2] University of Leicester, Leicester, UK

*Competing Interests:* No competing interests were disclosed.

**I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

---

**Version 1**

Referee Report 10 October 2018

https://doi.org/10.21956/wellcomeopenres.16048.r33935

✔ **Julian Wei-Tze Tang** [1,2]

[1] Clinical Microbiology, University Hospitals of Leicester NHS Trust, Leicester, UK
[2] University of Leicester, Leicester, UK

A very detailed, comprehensive write-up from the NGS methods viewpoint – which seems to be the main focus of this paper/studentship.

But a more global contextualization would make the paper more interesting and relevant to an international readership – this is very brief at present:

"Overall, the geographical distribution of sequences available was noted to be heavily biased towards the northern hemisphere and the Americas. This highlights the necessity to conduct further epidemiological surveillance of HPIV3 as well as the utility of a smaller part of the genome that could be used for this purpose, as identified in this study."

Especially in the Discussion where the UK sequences cluster with the other North/South American and European sequences. Why is there no clustering with South/East Asian PIV3 sequences? Were they too few or just not included in the alignment – or some other reason? See:

Godoy *et al.* (2016)[1]
Kim *et al*. (2017)[2]
Almajhdi *et al.* (2012)[3]
Piralla *et al.* (2009)[4]
Zhang *et al.* (2013)[5]

The authors should also discuss the possible impact of cell-cultured PIV3 viruses and how this may impact on the sequence fidelity in the resulting sequence obtained. This has been a problem for influenza viruses cultured in MDCK cell-lines for sequencing – and probably applies to other respiratory viruses cultured/amplified for sequencing:

Lee *et al.* (2013)[6]

Also as well as just analyzing the whole genome sequence phylogenetically, the authors should also analyse the individual PIV gene sequences ( at least where sufficient numbers of comparative sequences are available on Genbank) to see if the same cluster pattern is maintained compared to the whole genome phylogenetic analysis they have performed already. If the individual gene sequence analyses show different topologies, this may indicate some increased variability/ hypermutation/
recombination ongoing in some genes more than others, e.g. the fusion F gene. This
data may impact on vaccine design considerations, where a globally conserved region is an important target consideration - see:

Tang *et al.* (2017)[7]

**References**
1. Godoy C, Peremiquel-Trillas P, Andrés C, Gimferrer L, Uriona SM, Codina MG, Armadans L, Martín Mdel C, Fuentes F, Esperalba J, Campins M, Pumarola T, Antón A: A molecular epidemiological study of human parainfluenza virus type 3 at a tertiary university hospital during 2013-2015 in Catalonia, Spain. *Diagn Microbiol Infect Dis*. 2016; **86** (2): 153-9 PubMed Abstract | Publisher Full Text
2. Kim T, Jin CE, Sung H, Koo B, Park J, Kim SM, Kim JY, Chong YP, Lee SO, Choi SH, Kim YS, Woo JH, Lee JH, Lee JH, Lee KH, Shin Y, Kim SH: Molecular epidemiology and environmental contamination during an outbreak of parainfluenza virus 3 in a haematology ward.*J Hosp Infect*. 2017; **97** (4): 403-413

PubMed Abstract | Publisher Full Text

3. Almajhdi FN, Alshaman MS, Amer HM: Molecular characterization and phylogenetic analysis of human parainfluenza virus type 3 isolated from Saudi Arabia.*J Med Virol*. 2012; **84** (8): 1304-11 PubMed Abstract | Publisher Full Text

4. Piralla A, Percivalle E, Di Cesare-Merlone A, Locatelli F, Gerna G: Multicluster nosocomial outbreak of parainfluenza virus type 3 infection in a pediatric oncohematology unit: a phylogenetic study. *Haematologica*. 2009; **94** (6): 833-9 PubMed Abstract | Publisher Full Text

5. Zhang P, Ji YX, Xu ST, Yu DS, Yu PB, Zhang H, Xu J, Cui AL, Wei HL, Xu WB: [Genetic characterization of human parainfluenza virus 3 circulating in Gansu and Shaanxi Provinces from 2009 to 2011].*Bing Du Xue Bao*. 2013; **29** (5): 509-14 PubMed Abstract

6. Lee HK, Tang JW, Kong DH, Loh TP, Chiang DK, Lam TT, Koay ES: Comparison of mutation patterns in full-genome A/H3N2 influenza sequences obtained directly from clinical samples and the same samples after a single MDCK passage.*PLoS One*. 2013; **8** (11): e79252 PubMed Abstract | Publisher Full Text

7. Tang JW, Lam TT, Zaraket H, Lipkin WI, Drews SJ, Hatchette TF, Heraud JM, Koopmans MP, INSPIRE investigators: Global epidemiology of non-influenza RNA respiratory viruses: data gaps and a growing need for surveillance.*Lancet Infect Dis*. 2017; **17** (10): e320-e326 PubMed Abstract | Publisher Full Text

**Is the work clearly and accurately presented and does it cite the current literature?**
Partly

**Is the study design appropriate and is the work technically sound?**
Yes

**Are sufficient details of methods and analysis provided to allow replication by others?**
Yes

**If applicable, is the statistical analysis and its interpretation appropriate?**
Not applicable

**Are all the source data underlying the results available to ensure full reproducibility?**
Yes

**Are the conclusions drawn adequately supported by the results?**
Partly

*Competing Interests:* No competing interests were disclosed.

*Referee Expertise:* Clinical aspects, transmission and infection control, molecular epidemiology of influenza and other respiratory viruses.

**I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

Author Response 21 Nov 2018
**Anna Smielewska**, University of Cambridge Addenbrooke's Hospital Cambridge, UK

We would like to thank Dt Tang for his time and expertise and for the helpful comments. These have been addressed in version 2 of the manuscript as follows:

1. A note on geographical clustering has been added in Discussion
2. A note on the effects of cell culture adaptation has been added in Discussion
3. Analysis of the HN coding region has been included – additional figure (Figure 7) added in Results and a paragraph added in Discussion
4. Analysis using the F coding region was attempted, however this did not yield clustering with high enough bootstrap values and therefore was deemed to be an inadequate method for the current data set. A note to this effect has been added in Results.

***Competing Interests:*** none

Referee Report 03 October 2018

https://doi.org/10.21956/wellcomeopenres.16048.r33936

**Hirokazu Kimura** [1,2]

[1] School of Medical Technology, Faculty of Health Sciences, Gunma Paz University, Gunma, Japan
[2] Graduate School of Health Science, Gunma Paz University, Gunma, Japan

The authors performed detailed genetic analyses of human respirovirus 3 (HRV3) strains detected in UK using primer walking combined with the NGS method. I feel that methodology is sound. Moreover, the present genetic and molecular epidemiological data may be partly contribute to understand HRV3 virology. However, some major concerns were found.

1. HPIV3 has been reclassified into another Genus (Respirovirus) and species (respirovirus 3) by ICTV. It should be amended.

2. The authors phylogenetically analyzed using full genome of the present strains. Why?

3. No discussion of the phylogenetic relationships between your country and other countries was seen[1-2]. The authors could refer to them.

4. The present strains could be genotyped and/or subgenotyped as previously described[3].

**References**
1. Mizuta K, Tsukagoshi H, Ikeda T, Aoki Y, Abiko C, Itagaki T, Nagano M, Noda M, Kimura H: Molecular evolution of the haemagglutinin-neuraminidase gene in human parainfluenza virus type 3 isolates from children with acute respiratory illness in Yamagata prefecture, Japan.*J Med Microbiol*. 2014; **63** (Pt 4): 570-7 PubMed Abstract | Publisher Full Text
2. Takahashi M, Nagasawa K, Saito K, Maisawa SI, Fujita K, Murakami K, Kuroda M, Ryo A, Kimura H: Detailed genetic analyses of the HN gene in human respirovirus 3 detected in children with acute respiratory illness in the Iwate Prefecture, Japan.*Infect Genet Evol*. 2018; **59**: 155-162 PubMed Abstract | Publisher Full Text

3. Almajhdi FN: Hemagglutinin-neuraminidase gene sequence-based reclassification of human parainfluenza virus 3 variants.*Intervirology*. 2015; **58** (1): 35-40 PubMed Abstract | Publisher Full Text

**Is the work clearly and accurately presented and does it cite the current literature?**
Partly

**Is the study design appropriate and is the work technically sound?**
Partly

**Are sufficient details of methods and analysis provided to allow replication by others?**
Yes

**If applicable, is the statistical analysis and its interpretation appropriate?**
Not applicable

**Are all the source data underlying the results available to ensure full reproducibility?**
Yes

**Are the conclusions drawn adequately supported by the results?**
Partly

*Competing Interests:* No competing interests were disclosed.

**I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.**

Author Response 21 Nov 2018

**Anna Smielewska**, University of Cambridge Addenbrooke's Hospital Cambridge, UK

We would like to thank Dr Kimura for his time and expertise and for the helpful comments. These have been addressed in version 2 of the manuscript as follows:

1. A note on renaming of human parainfluenza virus 3 as human respirovirus 3 has been added in Introduction.
2. Additional clarification on why whole genome sequencing was used was added in Introduction
3. A note on geographical clustering has been added in Discussion
4. Analysis of the HN coding region was carried out, an additional figure (Figure 7) was added in Results and a paragraph added in Discussion.

*Competing Interests:* none