SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:

HIGH-THROUGHPUT SCREENING

PHYLOGENETICS

EVOLUTIONARY BIOLOGY

METAGENOMICS

Received
27 March 2013

Accepted
26 July 2013

Published
3 September 2013

Correspondence and requests for materials should be addressed to L.P.N. (lpn@ssi.dk)

Discovery of a divergent HPIV4 from respiratory secretions using second and third generation metagenomic sequencing

David E. Alquezar-Planas[1,2], Tobias Mourier[1], Christian A. W. Bruhn[1], Anders J. Hansen[1], Sarah Nathalie Vitcetz[1], Søren Mørk[3], Jan Gorodkin[3], Hanne Abel Nielsen[4], Yan Guo[5], Anand Sethuraman[5], Ellen E. Paxinos[5], Tongling Shan[6,7], Eric L. Delwart[7,8] & Lars P. Nielsen[2,9,10]

[1]Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Øster Voldgade 5-7, 1350 Copenhagen, Denmark, [2]Department of Virology, Statens Serum Institut, Artillerivej 5, 2300 Copenhagen, Denmark, [3]Center for non-coding RNA in Technology and Health, Department of Veterinary Clinical and Animal Science, University of Copenhagen, Grønnegårdsvej 3, 1870 Frederiksberg C, Denmark, [4]Department of Anesthesiology, Gentofte Hospital, Denmark, [5]Pacific Biosciences, Menlo Park, California, USA, [6]Department of Swine Infectious Disease, Shanghai Veterinary Research Institute (SHVRI), Chinese Academy of Agricultural Sciences (CAAS), [7]Blood Systems Research Institute, San Francisco, California, [8]Department of Laboratory Medicine, University of California at San Francisco, San Francisco, California, [9]Department of Clinical Microbiology, Odense University Hospital, Denmark, [10]Aalborg University, Department of Health Sciences, Aalborg, Denmark.

Molecular detection of viruses has been aided by high-throughput sequencing, permitting the genomic characterization of emerging strains. In this study, we comprehensively screened 500 respiratory secretions from children with upper and/or lower respiratory tract infections for viral pathogens. The viruses detected are described, including a divergent human parainfluenza virus type 4 from GS FLX pyrosequencing of 92 specimens. Complete full-genome characterization of the virus followed, using Single Molecule, Real-Time (SMRT®) sequencing. Subsequent "primer walking" combined with Sanger sequencing validated the *RS* platform's utility in viral sequencing from complex clinical samples. Comparative genomics reveals the divergent strain clusters with the only completely sequenced HPIV4a subtype. However, it also exhibits various structural features present in one of the HPIV4b reference strains, opening questions regarding their lifecycle and evolutionary relationships among these viruses. Clinical data from patients infected with the strain, as well as viral prevalence estimates using real-time PCR, is also described.

Acute respiratory viral infections are among the most common infections in humans. Most provoke only mild symptoms, but in some circumstances they may result in significant morbidity and mortality. In particular, infections with novel viruses can be severe when the host lacks preexisting immunity, conferred by prior encounters with the same or similar pathogens[1]. Not surprisingly, symptomatic respiratory tract infections are notably frequent in infants and young children, making them the most common cause of hospitalization (in more developed countries) for patients below the age of five[2].

The human parainfluenza viruses (HPIVs) are a group of four distinct serotypes of enveloped, non-segmented, single stranded negative sense RNA viruses belonging to the *Paramyxoviridae*. They are a significant cause of acute upper and lower respiratory tract infections in infants, frequently resulting in the need for in-patient care[3]. While the epidemiology and clinical manifestations of HPIV1-3 is well characterized, much less is known about the 4th serotype[4]. HPIV type 4 (HPIV4) was first identified in 1959 from a male college student exhibiting a mild upper respiratory tract infection[5], and has since been further divided into two distinct subtypes, 4a and 4b, based on antigenic differences[6]. Compared to other HPIVs, HPIV4 has been infrequently isolated in cell culture due to difficulties in propagation and a lack of cytopathic effects in most cell lines[7,8]. Moreover, the serotype has conventionally been reported to display milder clinical symptoms, which has led to its exclusion from routine diagnostic screening in most virology labs. By contrast, a number of studies have identified HPIV4 as a significant cause of respiratory disease especially in children[9–11]. Infections in immunocompromised patients[11] and in an otherwise healthy adult[12], resulting in acute respiratory failure, have been described. Recent epidemiological studies have shown that the clinical manifestations of HPIV4 resembled that of other HPIVs[13,14]. Moreover, accumulating evidence from a number of studies has identified HPIV4 as being more common than other serotypes[8,13,15,16]. In summary, these studies suggest that both the prevalence and clinical importance of HPIV4 as a major cause of respiratory illness, especially in co-infections with other viruses, may have been underestimated and remains poorly understood.

While specific and sensitive diagnosis of acute respiratory infections is important, in approximately 5-40% of cases (depending on the season) no infectious agent is identified[17]. The application of new sequencing methods to pathogen discovery can provide timely characterizations of novel viruses from various pathological tissues[18,19]. In this study, we comprehensively analyzed 500 respiratory samples from children with upper or lower respiratory tract infections for viral pathogens using a range of diagnostic techniques, including immunological, and PCR-based assays as well as Roche Genome Sequencer (GS) FLX Titanium pyrosequencing (Roche, Basel Switzerland) of 92 specimens. The identification of a novel variant (from a known viral pathogen) led us to further investigate its prevalence in these samples, and to characterize its genome by Single Molecule Real-Time (SMRT®) sequencing of the original clinical specimen on the PacBio® *RS* (PacBio) (Pacific Biosciences, Menlo Park, CA). We additionally explore the feasibility of single molecule sequencing, as well as verify the quality of the sequence obtained through this method to the complete Sanger sequenced viral genome.

Results

Immunological and molecular screenings of respiratory samples. Five hundred consecutive nasopharyngeal or tracheal aspirates from children admitted to Odense University Hospital due to upper and/or lower respiratory tract infection during the winter period of 2002 - 2003 were procured and analyzed for viral and bacterial pathogens. Molecular and limited epidemiological data from these screenings is reported in the supplementary results section, supplementary figure S1 and supplementary tables S1, S2, S3, S4, and S5.

The Identification and prevalence of a novel human parainfluenza virus type 4. An initial screening experiment of 92 pooled clinical samples on the GS FLX Titanium platform resulted in a total of six contigs and 39 singelton reads from one particular sample (459) with a significant local BLASTn-mediated GenBank match of 87-96% identity to HPIV4a. Presently, four complete genomes have been sequenced across both HPIV4 subtypes (one HPIV4a and three HPIV4b)[20–22]. Given that relatively little is known about these viruses we set out to recover the full genome in this sample (described below). We provisionally refer to this sequence as strain HPIV4_DK(459).

The exclusion of routine HPIV4 screenings in the public health setting in Denmark prompted us to also test the prevalence of this virus in 493 respiratory specimens (7 samples of the original 500 could not be located) using reverse transcriptase (RT) real-time PCR. PCR primers were designed from the HPIV4_DK(459) sequences discovered during GS FLX Titanium sequencing. The primers amplify a 61-bp fragment of the hemagglutinin-neuraminidase (HN) gene (see Material and Methods and table S6). PCR amplification was observed in a total of 13 samples (2.6%). Through an epidemiological perspective, the prevalence of HPIV4 described in this study corresponds to similar prevalence estimates to that of hMPV, Adenovirus, and *B. pertussis* using PCR-based screenings (see Supplementary Table S1).

Full genome of HPIV4_DK(459). Following the GS FLX screening and analysis, we sequenced sample 459 on the PacBio *RS* to recover HPIV4_DK(459) through a viral metagenomic framework. A total of 3,356,192 reads was generated from 5 SMRTcells® of sequencing. Viral read filtration of HPIV4 was achieved by mapping all sequences using Pacific Biosciences BLASR software to five HPIV reference genomes (x1 HPIV1, x1 HPIV2, x1 HPIV3, x1 HPIV4a, x1 HPIV4b). This resulted in 116,109 reads (~3.5%) mapping to the virus. Contigs were generated from filtered reads by *de novo* assembly, resulting in an assembly of 29 contigs, which were subsequently taxonomically identified through BLASTn (NCBI's non-redundant database). Despite sequence-independent amplification in the viral metagenomic preparation resulting in predominately short fragments as input (100–500 bp), up to 98% of the genome was covered at 700X average coverage from the contigs alone. A comparison of the coverage obtained from viral metagenomic sequencing of HPIV4_DK(459), (GS FLX and PacBio data) can be seen in Figure 1. We later confirmed the full genome sequence of HPIV4_DK(459) using conventional primer walking RT-PCR and 1st generation Sanger sequencing from GS FLX data output (GenBank accession number KF483663). The extremities (3′ and 5′ ends) of the genome were acquired in similar fashion using a consensus alignment and degenerate PCR amplification and Sanger sequencing. A list of the primers used for primer walking is shown in Supplementary Table S7. Using a simple majority-rule consensus approach, all positions with more than 100X PacBio coverage were identical to the Sanger sequenced genome, supporting the usefulness of this technique.

HPIV4_DK(459) genome analysis. The recovery and subsequent alignment of the divergent Danish strain confirmed a similar genomic orientation and content as found in the four other sequenced HPIV4 genomes. The genome spans 17,098 nucleotides, 46-nt longer than the only reported complete HPIV4a genome and more than 200-nt shorter than all three complete HPIV4b genomes

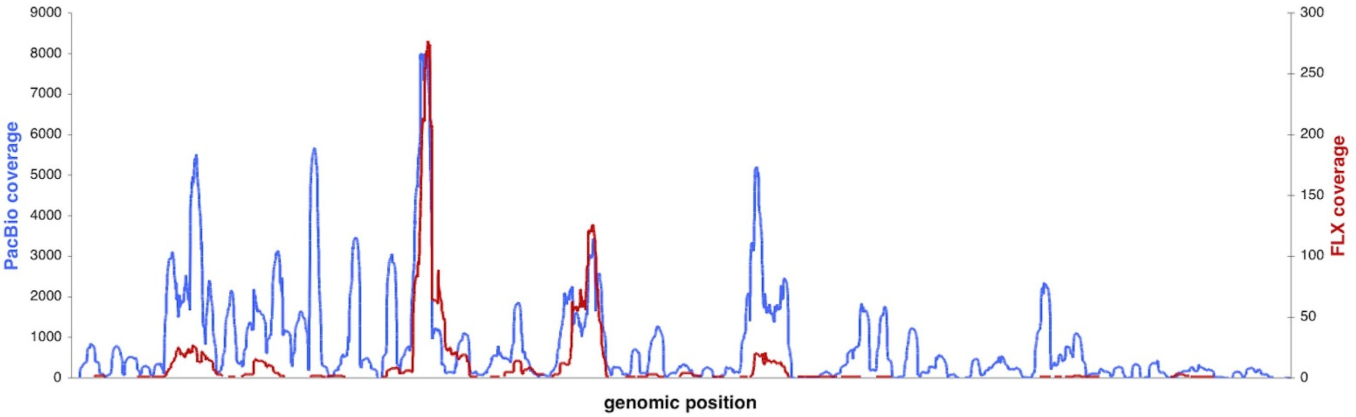

Figure 1 | Coverage plot of HPIV4_DK(459) (GenBank accession no. KF483663) from PacBio (blue line, left y-axis) and GS FLX reads (red line, right y-axis) compared to the genome sequence derived from Sanger sequencing.

Table 1 | A summary of the number of nucleotide changes between available HPIV4 complete genome sequences and HPIV4_DK(459) (GenBank accession no. KF483663)

	HPIV4 Structural Organization[1]												
HPIV Genome[2]	Int1	NP	Int2	P	Int3	M	Int4	F	Int5	HN	Int6	L	Int7
HPIV4a_AB543336.1													
Indels	57		1		3				29	17	17	2	2
Ist codon pos subst		13		12		2		7		20		42	
2nd codon pos subst		11		13				3		18		16	
3rd codon pos subst		45				21		48		68		188	
Total subst	10	69	25	53	34	23	68	58	92	106	95	246	19
HPIV4b_SKPIV4_EU627591.1													
Indels	3								8				272
Ist codon pos subst		28		33		19		29		46		103	
2nd codon pos subst		23		45		9		18		34		33	
3rd codon pos subst		139		82		113		122		170		539	
Total subst	12	190	85	160	100	141	141	169	200	250	249	675	52
HPIV4b_(68-333)_AB543337.1													
Indels	57		1		3				7	17	17		270
Ist codon pos subst		29		31		16		25		43		104	
2nd codon pos subst		22		46		18		13		32		35	
3rd codon pos subst		131		82		95		120		164		492	
Total subst	13	182	85	159	89	119	124	158	201	239	232	631	54
HPIV4b_(04-13)_JQ241176.1													
Indels	57		1		3				7	17	17		270
Ist codon pos subst		30		30		18		26		47		99	
2nd codon pos subst		23		47		8		16		32		32	
3rd codon pos subst		135		81		98		125		163		520	
Total subst	14	188	88	158	92	124	132	167	200	242	233	651	54

[1] Nucleotide differences are split into intergenic regions and six ORF's starting from 3′-NP-P-M-F-HN-L-5′.

[2] Abbreviations: 1st/2nd/3rd codon pos subst - 1st, 2nd or 3rd codon position substitutions, Total subst - Total substitutions, Intergenic region (Int), Nucleoprotein gene (NP), Phosphoprotein gene (P), Matrix gene (M), Fusion gene (F), Hemagluttanin-Neuraminidase gene (HN), Long gene (L).

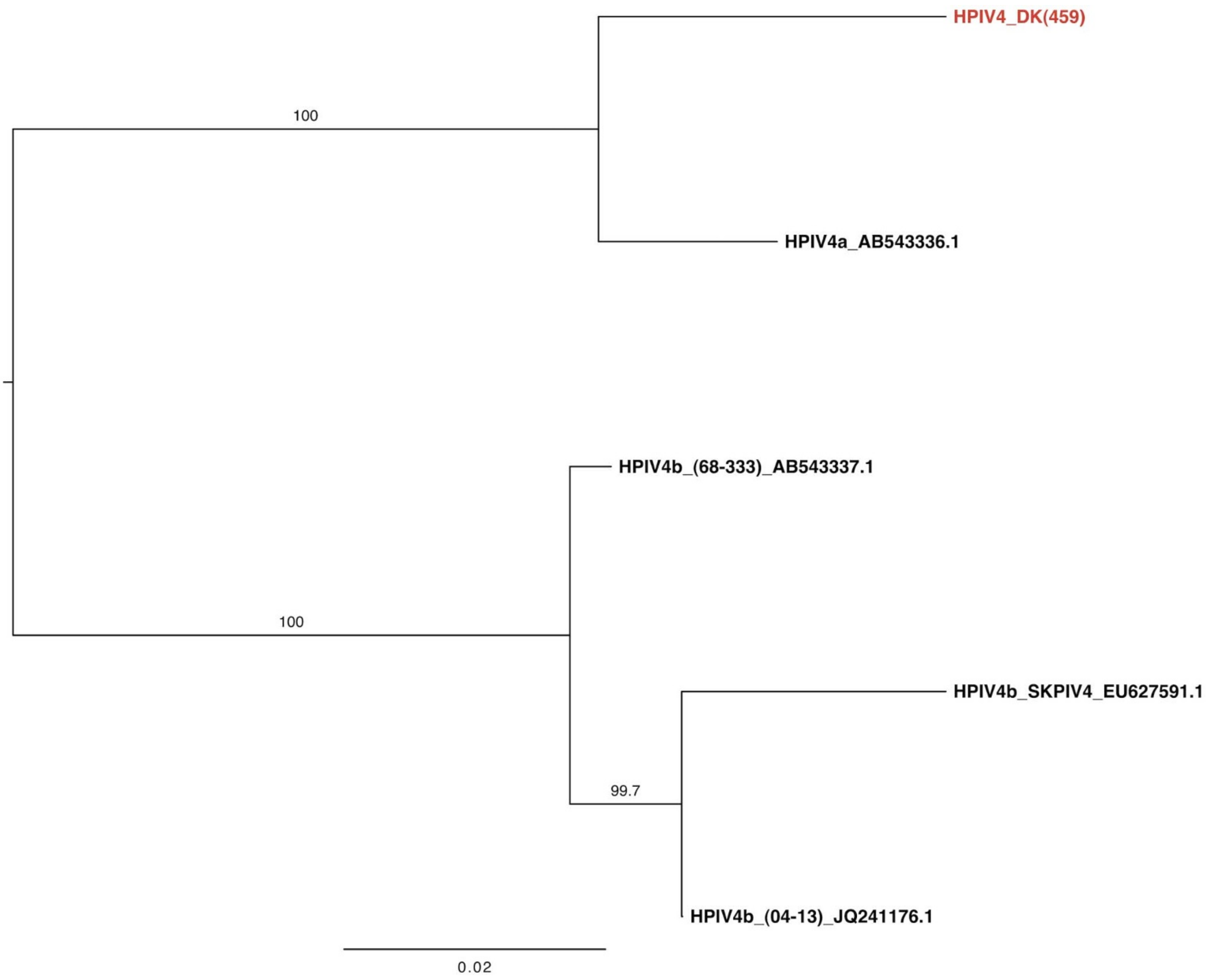

Figure 2 | Maximum likelihood tree from all 7 concatenated open reading frames in human parainfluenza virus 4. The tree was generated using the Tamura-Nei (TrN) nucleotide substitution model with 1000 bootstrap replicates, and midpoint rooted for clarity. Scale bar represents estimated phylogenetic distance in substitutions per site. HPIV4 strains are labeled with strain name and GenBank accession number on the right. The reported Danish strain, HPIV4_DK(459) (GenBank accession no. KF483663), is shown in red boldface.

described to date. An initial comparison of nucleotide changes across aligned HPIV4 genomes suggests that HPIV4_DK(459) appears more closely related to the 4a subtype than any of the 4b subtype strains (see Table 1). This is confirmed by phylogenetic analysis of an alignment of the concatenated coding regions (Figure 2) as well as of alignments of the entire genome (Supplementary Figure S2), and each gene separately (Supplementary Figures S3-9) making it the second HPIV4a genome sequenced to date.

The L gene in the *Paromyxaviridae* encodes the catalytic subunit of the RNA-dependent RNA polymerase protein used to transcribe and replicate the HPIV genome[3]. As this gene is likely to be relatively conserved, we chose it for inference of phylogenetic relationship to other viruses within the *Paramyxoviridae* using amino-acid sequences (see Figure 3). The analyses included members across seven genera including, the *Avulavirus, Henipavirus, Morbillivirus, Respirovirus Rubulavirus, Pneumovirus and Metapneumovirus*. As expected, tree topology clustered HPIV_DK(459) with all other HPIV4 genomes. Other members of the *Rubulavirus* genus, HPIV2 and mumps, formed a sister clade to HPIV4. The overall topology is congruent with previously published results[23].

Comparative genomics. Although HPIV4_DK(459) clusters most closely with HPIV4a than HPIV4b in all phylogenies, it shares four distinct features with the HPIV4b isolate SKPIV4[20], which are absent in all other completely sequenced HPIV4 genomes. Firstly, the isolate does not obey the 'rule of six' (the number of nucleotides in the genome being a multiple of six), a feature common for most but not all paramyxoviruses[24], which is thought to confer the ability of the RNA polymerase to increase genome replication[25].

Second, a 57-nt sequence-section is present at the 3′ leader intergenic region upstream of the NP gene. This 57-nt sequence is identical with the corresponding SKPIV4 sequence in all but 3 nucleotides. An initial prediction of this section's RNA folding structure using mfold[26] demonstrates an impressive stem-loop configuration, which folds into the exact same secondary structure (Supplementary Figure S10) as the SKPIV4 isolate, despite the minor differences at the nucleotide level. However, since single sequence secondary structure alone is generally not statistically significant for determining the presence of a structured RNA[27], we undertook a more elaborate analysis incorporating multi-sequence comparison and structural RNA alignments. We extended the analysis to include 17 additional

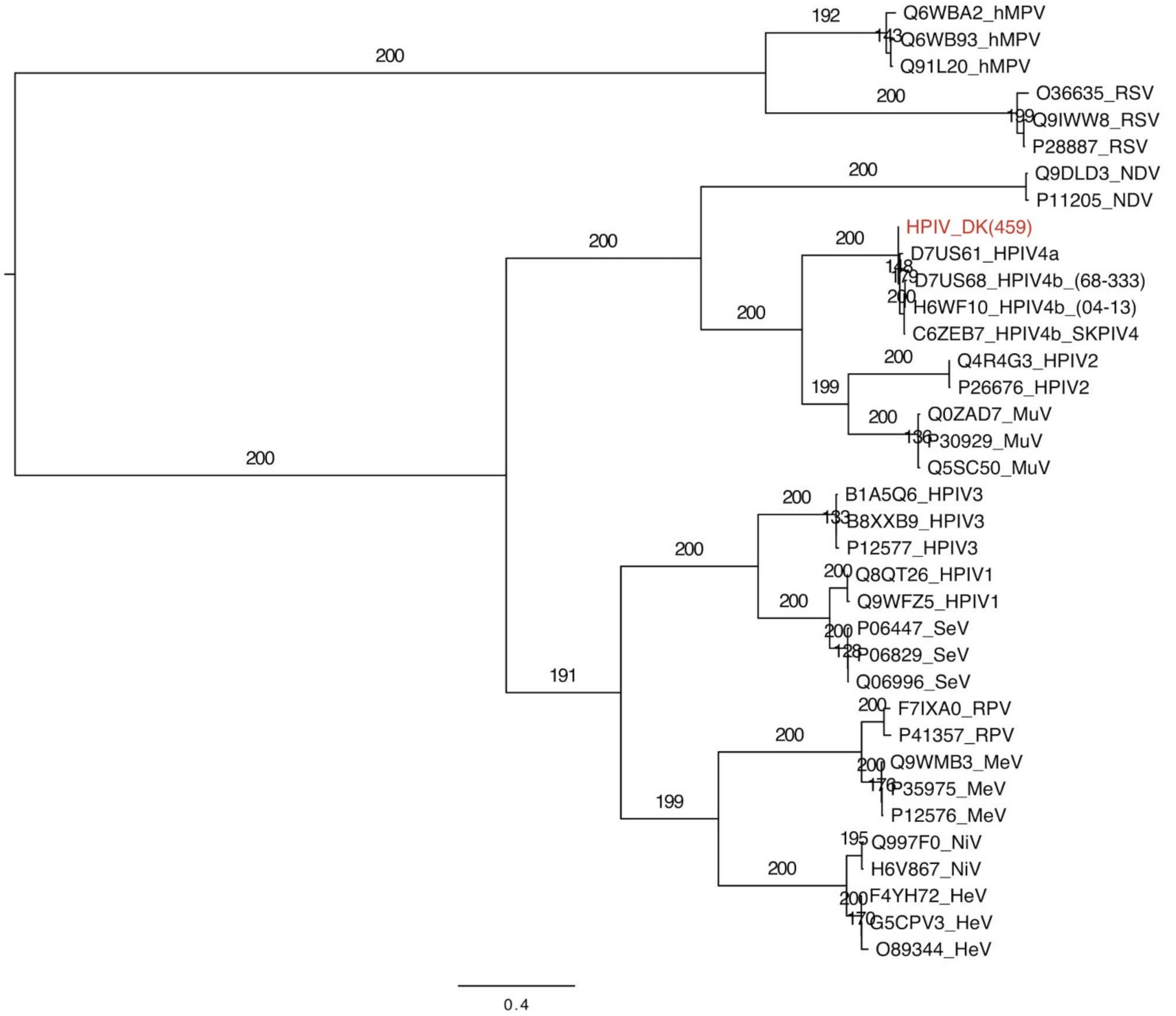

Figure 3 | Maximum likelihood phylogenetic tree of the L protein within members of the *Paramyxoviridae*. The tree was generated using the LG + G + F amino acid distance matrix with 200 bootstraps replicates and midpoint rooted for clarity. Branch length is proportional to estimated phylogenetic distance in amino acid substitutions per site. All viruses are labeled with respective L protein identifier from UniProt, followed by virus abbreviation as follows: human metapneumovirus (hMPV), respiratory synctial virus (RSV), Newcastle disease virus (NDV), human parainfluenza 1-4 (HPIV1-4), Mumps virus (MuV), Sendai virus (SeV), Rinderpest virus (RPV), Measles virus (MeV), Nipah virus (NV), Hendra virus (HeV). HPIV4 strains are additionally labeled with strain name. The reported Danish strain, HPIV4_DK(459) (GenBank accession no. KF483663) is shown in red.

Paramyxoviridae 3' proximal sequences from the genera Rubulavirus (HPIV2, HPIV4, Mumps, Simian virus 41, PIV5, Achimota 1, Tohuko 2, Menangle, Tioman), Henipavirus (Nipah virus), Morbillivirus (Measles), Respirovirus (HPIV1, HPIV3), and Avulavirus (Newcastle disease). A primary sequence alignment (Supplementary Figure S11) shows that taxa can be divided into two groups - those with ~ 100 nts 3' proximal sequences and those with ~ 150 nts 3' proximal sequences. The insert appears to be present in some Rubulaviruses whereas it is absent in the other genera. In order to estimate the reliability of the potential RNA structure we performed a comparative RNA structure analysis with a number of Rubulaviruses with 3' intergenic regions of similar lengths to the HPIV4_DK(459)

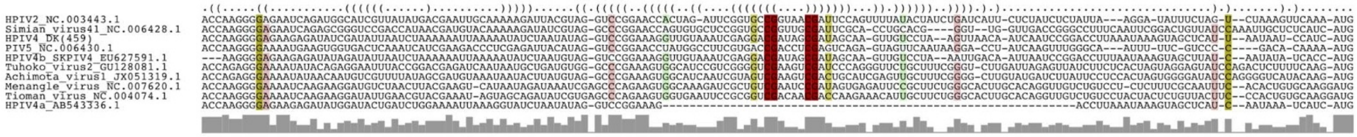

Figure 4 | Structural alignment of 150nt 3′ proximal sequences from Rubulaviruses produced with Stral and PETfold. A 100 nt HPIV4 sequence is overlaid showing the position of the insert with respect to the structure. Brackets refer to intra-molecular base-pairs of the RNA secondary structure. The extent of compensatory base changes is indicated using the Vienna RNA conservation coloring scheme[31]. The color indicates what type of base pair is formed (e.g. CG, AT, GU) and the strength of the color indicates the degree of conservation of the compensatory base change. The central stem structure is supported by the consensus structure of CMfinder, locarna, mxcarna, MASTR and PETfold[29]. The AUG at the end of each sequence is the start-codon of the NP gene. All viruses are labeled with virus name followed by GenBank accession number.

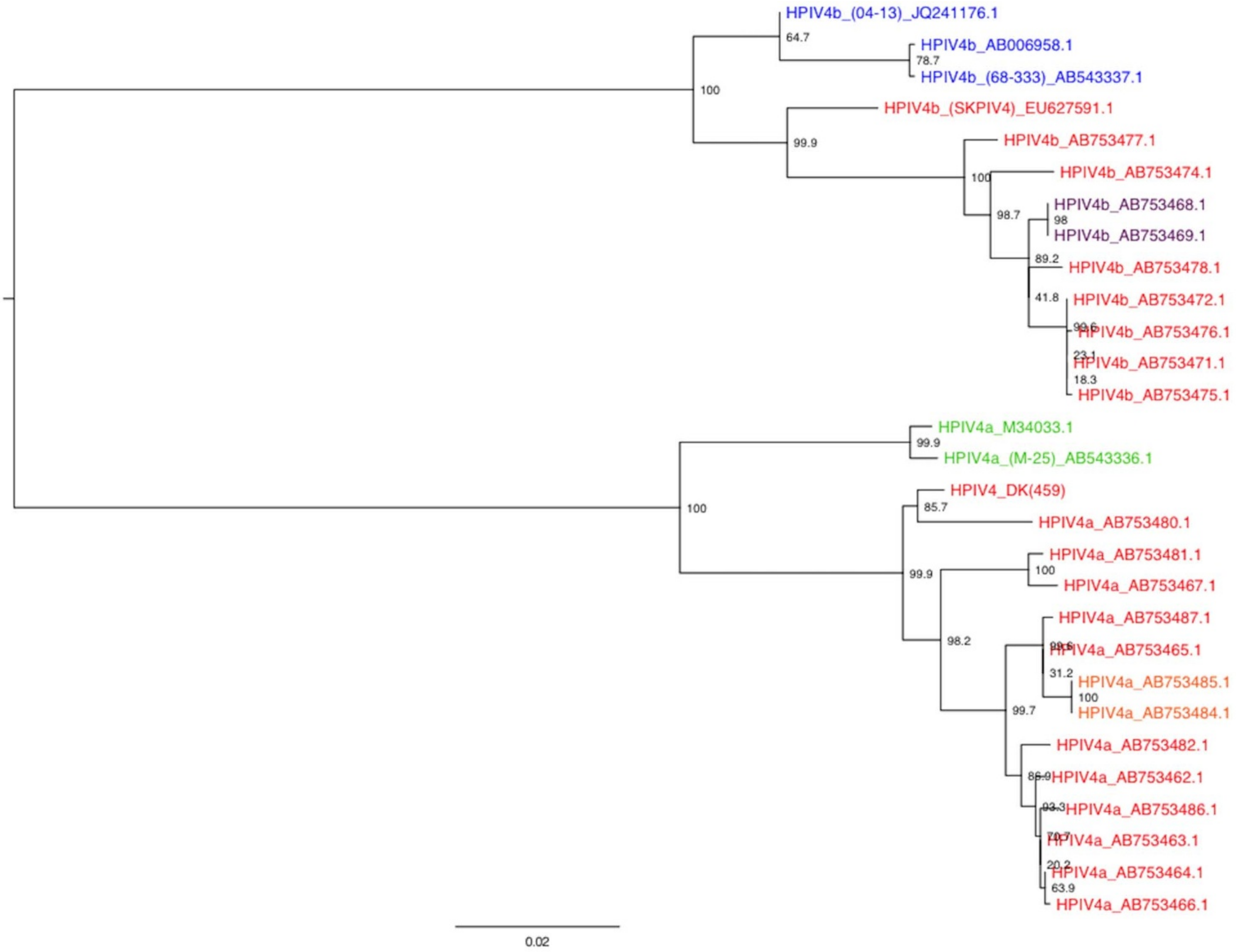

Figure 5 | Maximum likelihood tree from aligned hemagglutinin-neuraminidase gene in human parainfluenza virus 4. The presence of three indels in a section of approximately 60 nucleotides results in a frame-shift and changes at the amino acid level, generating five genotypes (one which is shared across both subtypes), highlighted in red (TQLLTYISYNGTI), orange (TQLFTYISYNGTI), purple (TQLLTYVSYNGTI) blue (LNWQHNYLHTYHIMVPL) and green (NNYLHIMVLSKI). The tree was generated using the HKY85 nucleotide substitution model with 1000 bootstrap replicates, and midpoint rooted for clarity. Scale bar represents estimated phylogenetic distance in substitutions per site. HPIV4 strains are labeled with strain name and GenBank accession number on the right of the sequence name. The reported Danish strain, HPIV4_DK(459) (GenBank accession no. KF483663), is found in red under the HPIV4a subtype.

sequence. Figure 4 shows the structural alignment produced using StrAl[28] and PETfold[29] via the WAR web server[30]. The central part of the conserved hairpin structure is also supported by MASTR, locarna, mxcarna, CMfinder and PETfold (http://genome.ku.dk/resources/war/). Using the sequenced genome of HPIV4_DK(459) we later designed a simple diagnostic real-time PCR assay (see Material and Methods and Supplementary Table S6) to test whether any other (n = 12) HPIV4 positive samples possessed the 57-nt sequence-section. Duplicate PCR amplification and subsequent Sanger sequencing confirmed that a minimum of 3 other samples (118, 294 & 309) contained the 'insert', suggesting that it is common among circulating Danish HPIV4 strains.

Thirdly, both genomes share an identical stretch of 13 amino-acid residues within the HN gene (TQLLTYISYNGTI), only one of which is conserved in all published complete HPIV4 genomes. Importantly, visual examination of aligned HN ORF nucleotide sequences (from all currently available HPIV4 sequences in GenBank [n = 29]) identified this stretch of amino acids in all but nine strains across both HPIV4 subtypes, which suggests a high prevalence among the HPIV4 strains. The differences observed in this section are due to three insertions-deletions (indels) across the strains that result in a frame shift at the amino acid level in this section. Based on this observation, up to five distinct genotypes can be visually identified from this data, three of which differ at only a single amino acid. This information is visually represented in Figure 5. An amino acid alignment of the HN gene from completely sequenced HPIV4 strains to two other paramyxoviruses (HPIV3, NDV) whose three-dimensional protein structure has been described using X-ray crystallography[32,33], places this stretch of 13 residues in the second Beta sheet, third strand (β_2S_3) of the globular head of the molecule. Meaningful inference of this sort is supported by the fact that the position of the protein structures of NDV and HPIV3 are highly conserved even though they are phylogenetically more distant, than either is to HPIV4. This structure is shown for HPIV3 and NDV to be immediately preceded by the α1 helix, which is seemingly crucial for the enzymatic activity of the protein[32]. Lastly, both HPIV4_DK(459) and SKPIV4 have a longer ORF at the C-terminal end of the HN gene, which codes for an extra five amino acids, the two first of which are identical. However, as above, (from a comparison of available HN sequences) the presence of the longer HN ORF appears more prevalent than its absence across strains. Using the same paramyxovirus alignment above, the C-terminal end of the HN gene can be inferred to reside in the end of the 6th Beta sheet, third strand (β_6S_3) of the globular head of the protein molecule. This beta sheet has been identified as having catalytic/active sites in both the first and second strand[32]. The amino acid residue differences within the HN gene of each of the five sequenced HPIV4 reference genomes can be visualized in Supplementary Figure S12.

Discussion

Advances in molecular technologies have allowed ever more rapid and sensitive identification of etiological agents of infectious diseases, including epidemics. In clinical virology, PCR based assays still remain the gold standard. However, their high specificity and the emergence of divergent strains may cause the assay to fail routine diagnosis, leading to elevated levels of false negatives. High-throughput metagenomic sequencing provides an alternative measure, by supplying added sequence information and revealing details on the divergence to characterized strains. Here we show that not only 2nd generation sequencing platforms, but also 3rd generation single molecule sequencing may be used effectively in viral discovery. Although one of the primary strengths of the PacBio lies in the generation of long sequences, the platform can clearly be used for short read sequencing. The pre-sequencing amplification methods used in this study primarily generated short input fragments. However, using few SMRT cells we obtained excellent coverage of the genome. Given its fast turnaround time, the PacBio is well suited for rapid and relatively cheap characterization of novel viruses. In particular, when other more sensitive screening (sequencing) platforms have identified candidate viral reads within the clinical samples tested.

Amongst the identified viruses using deep sequencing was the presence of a divergent HPIV4. HPIV4 has been traditionally described as fastidious in nature, and less clinically important than other HPIV serotypes[4]. Overall this has resulted in few epidemiological and case report studies being published[34], and its infrequent screening in clinical microbiology laboratories[11]. In temperate climates, peak incidence of HPIV4 is often reported during the late fall and late winter months[4,14,16,34]. However, other studies have detected peaks outside the usual seasons[13,34], emphasizing a poorly understood epidemiology and seasonality[35]. Aside from this consideration, several other factors may influence its detection including geographical location, sample population, and the year of collection[8,11]. Although the identified proportion of HPIV4 positive samples in our study was within previously described ranges, the aforementioned factors will likely result in significant differences across spatial and temporal distributions. More importantly, the lack of complete genomes confirms how little is known of circulating HPIV4 viral strains, emphasizing the need for focused full genome studies. In accordance with the above, we recovered the full genome of a divergent HPIV4 in a proof of principle study using 3rd generation metagenomic sequencing. Despite the short lengths of the amplicons, confirmatory Sanger sequencing later attested to the validity of this approach in virus discovery.

All phylogenetic analyses convincingly grouped HPIV4_DK(459) with HPIV4a (see Figure 2 and Supplementary Figures S2-9) However, interestingly, strain HPIV4_DK(459) shares various features with one HPIV4b isolate (SKPIV4), which phylogentically otherwise appears the more distantly related to HPIV4_DK(459) in all but two analyses performed. This included the presence of a 57-nt sequence section at the 3′ leader intergenic region prior to the NP gene. Given that conserved promoter sequences in this region of PIV's is believed to drive transcription and replication of nascent RNA into postive-sense RNA replicas[24], it makes it alluring to speculate on the functional role of this section. Interestingly, a number of the *Paramyxoviridae* have 3' intergenic regions (up to the start codon of the NP gene) of around 150 nts, approximately the same length as the HPIV4_DK(459) strain and SKPIV4 sequences. In silico generated RNA structure analysis revealed that the only consistent consensus secondary structure overlaps the 57-nt 'insert' region (see Figure 4). These results indicate that the insert is potentially necessary for RNA secondary structure formation that seems to be conserved across different members of the *Paramyxoviridae*. The 3' leader sequence is known to be involved in RNA replication, transcriptional and encapsidation regulation[36]. The potential regulatory consequences of the putative RNA structure in the leader sequence remain to be examined.

Moreover, under the premise that the ends of the genome were suitably acquired, neither genome conforms to the rule of six, a feature thought to influence replication efficiency within the *Paramyxoviridae*[37]. Two additional features shared with the SKPIV4 strain are both located within the HN gene, which encodes one of two proteins expressed on the viral surface. The HN envelope glycoprotein has three important functions. First, it mediates cellular infection through binding of the virus to cell-surface receptors containing sialic acid. Second, it severs the binding of the cells receptor from the glyco-protein conjugates during viral budding. Finally, through interaction with a second surface glycoprotein, it regulates fusion between the virus envelope with the cells plasma membrane[24]. Here, we have provisionally inferred the location of these conserved residues by aligning the HN gene to two members of the *Paramyxoviridae* (NDV, HPIV3), whose crystallographic three dimensional protein structures have been described[32,33]. However, the biological and functional relevance, such as the mediation of cellular receptor binding or Fusion protein activation remains unclear. Previous studies have ascertained that both the stalk region and the globular head are functionally important[38], thus investigation of these mechanisms will require various directed studies.

While it will become important to clarify the functional role of the shared features between HPIV4_DK(459) and SKPIV4 (if any), it is equally intriguing to speculate on their evolutionary history. Although the differences described in our viral strain (within the HN gene) can be attributed to a number of indels, the presence of the 57-nt sequence-section cannot. The occurrence of this common sequence-section across HPIV4 subtypes may be explained through a common evolutionary history.

The high level of selective pressure on the HN gene by the host's immune system suggests that these features (the stretch of 13 amino acid sequence section and the ends of the HN genes) might be under strong positive selection, which is clearly supported by their positions in the putative protein structure. Additionally, this might also be the case for the RNA hairpin structure, which potentially regulates part of the viral life cycle. Indeed, studies of individual viral genera do indicate varying patterns of selection acting within different genes. However, testing such effects at different taxonomic hierarchical levels are likely to be problematic given low species sampling sizes[23].

The lack of genomic data and the neglected epidemiology of HPIV4 leave questions unanswered. Although our results link HPIV4_DK(459) to the SKPIV4 isolate, the implications of these conserved regions in the genome is presently unclear. Further resolution into the location of these conserved residues, perhaps beginning with crystallographic structural analysis of HN proteins or protein folding prediction, may shed light on their function. Moreover, having described the precise location on the molecules three-dimensional structure, site-directed mutational studies could further resolve their biological relevance[39]. With reduced costs in high-throughput sequencing future studies should additionally aim on recovering complete high quality genomes from various HPIV4 strains. Such an endeavor, will further resolve evolutionary relationships between homologous serotypes, HPIV4 subtypes, and sister species.

Methods

Respiratory specimens. Five hundred nasopharygeal aspirates or tracheal secretions previously collected from children during the winter season of 2002-2003 admitted to Odense University Hospital, Denmark were procured. All samples were taken from children of ages 0 to 5 years hospitalized for symptoms of acute respiratory infection. Laboratory and limited clinical data was collected and samples subsequently anonymized.

Immunological and RT-PCR based pathogen screening. All samples were initially tested for respiratory syncytial virus using either Immunofluorescence (Dako) or rapid immunological tests performed with membrane ELISA (Beckton-Dickenson). As only 36% of the samples collected during the RSV season tested positive, we tested for a range of pathogens (RSV, influenza A and B, human metapneumovirus, parainfluenza type 3, coronavirus OC43, NL-63 and 229E, rhinovirus, enterovirus, parechovirus, adenovirus, bocavirus, *Mycoplasma pneumonia*, *B. pertussis and C. pneumoniae*) using RT-PCR and PCR with amplicon hybridization to probe coated microtiter plates as described elsewhere[40–42]. From this cohort of 500 respiratory samples, 92 samples that tested negative after the aforementioned diagnostic tests were subjected to viral metagenomics to fill the diagnostic gap.

Viral particle purification and extraction for GS FLX sequencing. A volume of 500 µl from 92 respiratory aspirates were clarified by centrifugation at 12,000 rpm for 2 mins to pellet any cellular debris. The supernatant was then collected and passed through a 0.45 µM sterile filter (Millipore) at 12,000 rpm for 5 minutes for further removal of cellular material and bacterial sized particles. The filtrates were collected and spun in an ultracentrifuge (Beckman Coulter Optima LE-80) at 32,000 rpm for two hours at 8°C. Sample supernatant was discarded, and viral particles resuspended with 110 µl of Hanks buffered saline. Subsequently, the filtrates were treated with a cocktail of DNase and RNase enzymes that included Turbo DNase (Ambion), Baseline ZERO (Epicentre), Benzonase (Novagen), RNase A (Fermentas) in a x1 Turbo DNase reaction buffer to remove unprotected nucleic acids as previously described[18]. Viral nucleic acids were then extracted using the QIAamp viral RNA extraction kit (Qiagen) according to the manufacturers instructions.

Respiratory virome library construction for GS FLX sequencing. Nucleic acid extracts from each sample were treated separately in either 'RNA only' or an 'RNA and DNA' sequence-independent amplification route as described previously[18,43]. Briefly, 10 µl of extracted nucleic acids was incubated with DNase (Ambion) for an RNA virus-only enrichment. For the RNA and DNA enrichment the nuclease treatment step was omitted. Both reactions were then reverse transcribed using Superscript III (Invitrogen) and 100 µM of a distinct random primer consisting of a 20-bp nucleotide sequence at the 5' end and a randomized octamer sequence at the 3' end. Following reverse transcription, the cDNA was made double stranded by using a single round of Klenow fragment polymerase (New England Biolabs). PCR amplification was performed in duplicate for both the RNA and RNA plus DNA fractions by using primers consisting of only the 20-bp fixed portion of the random primer. This resulted in a total of four PCR reactions per sample, which was subsequently pooled and purified using the QIAquick PCR purification kit (Qiagen) and the DNA concentration determined by nanodrop (Thermo scientific). The amplified products were run on a 2% agarose gel, yielding a DNA smear. Fragments of approximately 500 to 1000-bp were excised from the gel, and purified using the QIAquick Gel Extraction Kit (Qiagen). Equimolar concentrations of amplified samples were pooled and built into two separate DNA libraries and run on the GS FLX titanium platform (454 Life Sciences).

GS FLX pyrosequencing read analysis. Sequence reads were filtered through sequence similarity to local human (human genome hg18) and ribosomal sequence databases using BLAT. Remaining filtered reads were then identified and binned according to the different PCR primers used to amplify the sample. Each read was subsequently trimmed of the primer consisting of a fixed 20-bp nucleotide sequence and an adjacent 8-bps representative of the random portion of the primer. All filtered reads were later mapped against the European Bioinformatics Institute's (EBI) viral and phage genome reference databases (http://www.ebi.ac.uk/) using ssaha2 for taxonomic identification[44]. Additionally, filtered binned reads were built into contigs using Newbler, and taxonomically identified by comparison to GenBank (best hit) as bacterial, viral or unknown using BLASTn and BLASTx searches. A cutoff off value of $\leq 10^{-5}$ was used for determination of the significance of a hit. Any sequence with an E value of $> 10^{-5}$ was deemed unclassifiable and removed.

Viral particle purification and extraction for PacBio RS sequencing. Complete genome sequencing of the divergent HPIV4 (sampe 459) was performed through a metagenomic framework and sequenced on the PacBIO RS. The viral particles were treated in a manner like the 454 sequencing experiment, but with a few notable exceptions. Three aliquots of 150 µl of respiratory sample was clarified and filtered in a 0.45 µM sterile filter (Costar) as described above. Nuclease treatment employed double the concentration previously described[18]. All other experimental processes up to and including the PCR product pooling and purification was performed alike the 454-processing-pipeline with the exception that 8 amplifications (x4 RNA only and x4 RNA plus DNA) were performed. Subsequently, the purified DNA products were measured using a Qubit fluorometer (Invitrogen), built into libraries and sequenced.

PacBio RS library construction and sequencing. Small fragments (<100 bp) in the DNA sample were removed prior to library construction with one purification of 1.8 × volumes of Agencourt AMPure XP beads (Beckman Coulter Genomics) as per the manufacturers instructions. Fragment size assessment and quantification was verified using a Bioanalyzer 2100 with DNA 7500 chemistry (Agilent). A single SMRTbell library was generated using standard reagents and protocols (overhang ligation protocol). As the input fragments are significantly smaller than those previously tested on the PacBio RS, we sequenced the library at 5 different on-chip concentrations (80 pM, 40 pM, 20 pM, 10 pM, 5 pM).

PacBio RS sequence analysis. Total sequence output from the five SMRTcells was then filtered by mapping against five HPIV reference genomes (HPIV1 - NC_003461.1, HPIV2 - NC_003443.1, HPIV3 - NC_001796.2, HPIV4a - AB543336.1, HPIV4b - EU627591.1) using BLASR. To increase the stringency of the bioinformatic approach, mapped HPIV4 reads were further filtered on a minimum pass accuracy of 80% and minimum read length of 100-bp. Assembly of filtered reads into contigs was subsequently achieved using the Allora software[45] then taxonomically identified using BLASTn by comparison to GenBanks non-redundant database.

Sanger sequencing. Primer walking and Sanger sequencing was additionally used to obtain the full genome of HPIV4_DK(459) and for confirmation of the amplified 57 nucleotide insert. All homologs (singlet reads and contigs from GS FLX data) matching HPIV4_DK(459) were aligned with the four currently available HPIV4 genomes in GenBank (HPIV4a - accession numbers AB543336.1, HPIV4b - accession numbers EU627591.1, AB543337.1, JQ241176.1) using Geneious v5.4.2 software[46]. Multiple primer sets were designed from the alignment, including consensus primers to fill the ends of the sequences and any significantly large gaps within the genome. Primers were designed using Primer3[47] and prepared by Eurofins MWG Operon, Ebersberg Germany (Supplementary Table S2). Amplification of each viral target was achieved by using a OneStep RT-PCR kit (Qiagen) in a 50 µl reaction volume. The final reaction mixture contained 1x OneStep RT-PCR buffer, 400 µM of each deoxynucleoside triphosphate (dNTP), 2.5 mM of $MgCl_2$, 0.6 µM of each primer, and 2 µl of OneStep RT PCR enzyme mix. PCR cycling conditions varied depending on the estimated size of the amplified product (Supplementary Table S2). Amplified material was sent to Macrogen (The Netherlands) for purification and Sanger sequencing in both directions by following the companies recommendations and use of the same PCR amplification primers for sequencing. Following full genome recovery, sequenced fragments were aligned to the four reference genomes and then assembled using Geneious v5.4.2[46].

To compare genome coverage between GS FLX and PacBio RS, unfiltered GS FLX and PacBio reads were mapped against the Sanger sequences obtained by Primer Walking using ssaha2[44]. Using the pileup feature in samtools[48], the number of reads with base calls divergent from the consensus were recorded and plotted along with coverage.

Real-time quantitative PCR (qPCR) of divergent HPIV4. *Nucleic acid extraction*. Viral RNA was extracted from respiratory secretions using the MagNa Pure LC Total Nucleic Acid Isolation Kit (Roche Diagnostics) on the KingFisher semi-automated magnetic particle processor (Thermo Scientific) following the manufacturers instructions. For quality control purposes, two negative template controls (one per plate) were simultaneously extracted in each KingFisher run, where sample was replaced with sterile water (Ambion). Nucleic acid extractions from each respiratory secretion or negative template control were then stored at -20°C in RNase free microfuge tubes.

Multiplexed real-time PCR primers, probes and positive controls. Aligned pyrosequencing data (described above) was used to design a multiplexed TaqMan qPCR assay for qualitative detection of a 61 bp fragment matching the HN gene in HPIV4_DK(459). All real-time primers and probes (Supplementary Table S6) were designed using Primer3 software[48] and prepared by Pentabase (Odense Denmark). To ensure analytical performance, a synthetic DNA control was designed consisting of the same primer and probe sites for the HPIV4 target (HN real-time assay not insert real-time PCR assay) and an additional 5 S ribosomal RNA sequence for distinction between real and false positives. The DNA oligonucleotide was ordered from Pentabase (Odense Denmark), (5′CAAACCCATTGGTGTTATACCCAGTAGTTTGATT-GCTAGATGCAGGGTCTGCACAGTTGTGCAGTGTCTGCACATCCTTGTAG-GAGTCGTGCGTCGATGAAGAACGCAAGTGCAATCAATATGTATCACTAG-GAACC3′).

Real-time PCR 'insert' assay. A real-time RT-PCR assay was also designed on the aligned genomes (using Geneious)[46] of all five completely sequenced HPIV4 strains. The primers were designed using the same software (as above) and prepared by TAG Copenhagen (Copenhagen Denmark).

Real-time PCR assays. All real-time RT-PCR assays used 5 µl of template nucleic acids from each respiratory sample to 20 µl of reaction mix. The reaction mix was made using the OneStep RT-PCR kit (Qiagen) and contained 5X OneStep reverse transcription (RT)-PCR buffer, 400 µM of each deoxynucleoside triphosphate (dNTP), 2.5 mM of $MgCl_2$, 0.5 mM of each primer, 0.1 mM of the TaqMan probe and 1 µl of OneStep RT PCR enzyme mix. PCR cycling consisted of an initial cDNA synthesis step at 50°C for 30 min followed by an initial denaturation step at 95°C for 15 min and then 45 cycles of amplification at 95°C for 15 s, 55°C for 60 s and 72°C for 15 s. Real-time amplification was performed on the Roche LC480 light cycler. Negative template controls and multiple diluted positive controls were simultaneously run in all (HN assay only) real-time PCR reactions.

Phylogenetic analysis. *Human parainfluenza viruses type 4 (HPIV4)*. Alignment of the full Sanger sequenced genome of isolate HPIV4_DK(459) and all the four other available HPIV4 reference genomes, as well as all coding regions, individually and concatenated, were achieved using the MAFFT plug-in[49] in Geneious[46] with subsequent visual inspection and manual correction. Signs of saturation on the third codon postion of the alignments were tested using DAMBE[50], but none were found. Nucleotide substitution models were chosen using the Bayesian Information criterion in ModelGenerator[51] and Maximum likelihood trees built using the PHYML plugin[52] in Geneious with 1000 bootstrap replicates. All trees were visually inspected and annotated using the FigTree graphical viewer interface version 1.3.1 (http://tree.bio.ed.ac.uk/software/figtree/) including mid-point rooting for clarity.

Paramyxoviruses. A phylogenetic relationship within the *Paramyxoviridae* was inferred using amino acid sequences from the L protein. Sequences were imported from the UniProt database, - including multiple paramyxoviridae strains (n = 31) in addition to the x5 HPIV4 sequences (see protein identifiers on tree) - and, three initial alignments constructed using three different scoring matrices (Blosum62, Jtt100, Jtt200) to compare the quality of the alignments. All matrix alignments were performed in Geneious v5.6.4[46] using the MAFFT v6.717 b plugin[49] and the standard open gap penalties, and offset values. The Blosum62 matrix produced the best quality alignment based on direct visual inspection and using Entropy-Two (www.hiv.-lanl.gov). This alignment was then refined using the Geneious Alignment software[46]. Amino-acid saturation was tested using ASaturA[53], and together with visual inspection this led to the removal of a significant number of sites including virtually all gap containing positions and their often ambiguously aligned neighboring positions, resulting in a final alignment, 1567 amino acid residues in length. ProtTest[54] was used to test the models of protein substitution; both the AIC, AICc and BIC criteria chose the LG (Le Gascuel) + G + F as the best fit[55] The online version of PhyML[52], (http://www.atgc-montpellier.fr/phyml/) was used to infer the Maximum Likelihood tree using 200 bootstrap replicates. The tree was then visualized and annotated using the FigTree graphical viewer interface version 1.3.1 (http://tree.bio.ed.ac.uk/software/figtree/) including mid-point rooting for clarity.

1. Hussell, T., Godlee, A., Salek-Ardakani, S. & Snelgrove, R. J. Respiratory viral infections: knowledge based therapeutics. *Curr. Opin Immunol.* **24**, 438–443 (2012).
2. Henrickson, K. J., Hoover, S., Kehl, K. S. & Hua, W. National disease burden of respiratory viruses detected in children by polymerase chain reaction. *Pediatr Infect Dis J.* **23**, S11–8 (2004).
3. Henrickson, K. J. Parainfluenza Viruses. *Clin Microbiol Rev.* **16**, 242–264 (2003).
4. Lau, S. & To, W. Human parainfluenza virus 4 outbreak and the role of diagnostic tests. *J Clin Microbiol.* **43**, 4515–4521 (2005).
5. Johnson, K. M., Chanock, R. M., Cook, M. K. & Huebner, R. J. Studies of a new haemadsorption virus.1. Isolation, properties and characterization. *Am J Hyg.* **71**, 81 (1960).
6. Canchola, J. G., Vargosko, A. J. & Kim, H. W. Antigenic variation among newly isolated strains of parainfluenza type 4 virus. *Am J Hyg.* **79**, 357-364 (1964).
7. Gardner, S. D. The isolation of parainfluenza 4 subtypes A and B in England and serological studies of their prevalence. *J Hyg.* **67**, 545–50 (1969).
8. Fairchok, M. P., Martin, E. T., Kuypers, J. & Englund, J. a. A prospective study of parainfluenza virus type 4 infections in children attending daycare. *Pediatr Infect Dis J.* **30**, 714–6 (2011).
9. Rubin, E. Infections due to parainfluenza virus type 4 in children. *Clin Infect Dis.* **17**, 998–1002 (1993).
10. Lindquist, S. W., Darnule, A., Istas, A. & Demmler, G. J. Parainfluenza virus type 4 in pediatric patients. *Pediatr Infect Dis J.* **16**, 34-38 (1997).
11. Lau, S. K. P. *et al.* Clinical and molecular epidemiology of human parainfluenza virus 4 infections in Hong Kong: subtype 4B as common as subtype 4A. *J Clin Microbiol.* **47**, 1549–52 (2009).
12. Chiu, C. Y. *et al.* Microarray detection of human parainfluenzavirus 4 infection associated with respiratory failure in an immunocompetent adult. *Clin infect Dis.* **43**, e71–6 (2006).
13. Ren, L. *et al.* Human parainfluenza virus type 4 infection in Chinese children with lower respiratory tract infections: a comparison study. *J Clin Virol.* **51**, 209–12 (2011).
14. Liu, W.-K. *et al.* Epidemiology and clinical presentation of the four human parainfluenza virus types. *BMC Infect Dis.* **13**, 28 (2013).
15. Aguilar, J. C. *et al.* Detection and identification of human parainfluenza viruses 1, 2, 3, and 4 in clinical samples of pediatric patients by multiplex reverse transcription-PCR. *J Clin Microbiol.* **38**, 1191–5 (2000).
16. Vachon, M. *et al.* Human Parainfluenza Type 4 Infection, Canada. *Emerg Infect Diseases.* **12**, 1755–1758 (2006).
17. Regamey, N. *et al.* Viral etiology of acute respiratory infections with cough in infancy: A community-based birth cohort study. *Pediatr Infect Dis J.* **27**, 100–105 (2008).
18. Victoria, J. G. *et al.* Metagenomic analyses of viruses in stool samples from children with acute flaccid paralysis. *J Virol.* **83**, 4642–51 (2009).
19. Daly, G. M. *et al.* A Viral discovery methodology for clinical biopsy samples utilising massively parallel next generation sequencing. *PLoS One.* **6**, e28879 (2011).
20. Yea, C. *et al.* The complete sequence of a human parainfluenzavirus 4 genome. *Viruses.* **1**, 26–41 (2009).
21. Komada, H. *et al.* Completion of the full-length genome sequence of human parainfluenza virus types 4A and 4B: sequence analysis of the large protein genes and gene start, intergenic and end sequences. *Arch Virol.* **156**, 161–6 (2011).
22. Lednicky, J., Waltzek, T., Halpern, M. & Hamilton, S. Comparative analysis of the full-length genome sequence of a clinical isolate of human parainfluenza virus 4B. *Scientifica.* 4–7 (2012).
23. McCarthy, A. J. & Goodman, S. J. Reassessing conflicting evolutionary histories of the Paramyxoviridae and the origins of respiroviruses with Bayesian multigene phylogenies. *Infect Genet Evol.* **10**, 97–107 (2010).
24. Psarras, S., Papadopoulos, N. G. & Johnston, S. L. Parainfluenza Viruses. *Principles and Practice of Clinical Virology* (Zuckerman, A., Banatvala, J., Schoub, B., Griffiths, P. & Mortimer, P.) 409–439 (John Wiley & Sons Ltd., 2005).
25. Kolakofsky, D., Pelet, T., Garcin, D., Curran, J. & Roux, L. Paramyxovirus RNA synthesis and the requirement for hexamer genome length: the rule of six revisited. *J Virol.* **72**, 891–899 (1998).
26. Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **31**, 3406–3415 (2003). Last accessed 6th June 2012.
27. Rivas, E. & Eddy, S. Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics.* **16**, 583–605 (2000).
28. Dalli, D., Wilm, A., Mainz, I. & Steger, G. STRAL: progressive alignment of non-coding RNA using base pairing probability vectors in quadratic time. *Bioinformatics.* **22**, 1593–9 (2006).
29. Seemann, S. E., Gorodkin, J. & Backofen, R. Unifying evolutionary and thermodynamic information for RNA folding of multiple alignments. *Nucleic Acids Res.* **36**, 6355–62 (2008).
30. Torarinsson, E. & Lindgreen, S. {WAR}: Webserver for aligning structural {RNAs}. *Nucleic Acids Res. Web Server Issue* W79–84 (2008). Last accessed 20th Mar 2013.
31. Lorenz, R. *et al.* ViennaRNA Package 2.0. *AMB.* **6**, 1–14 (2011).
32. Crennell, S., Takimoto, T., Portner, a. & Taylor, G. Crystal structure of the multifunctional paramyxovirus hemagglutinin-neuraminidase. *Nature Struct Biol.* **7**, 1068–74 (2000).
33. Lawrence, M. C. *et al.* Structure of the haemagglutinin-neuraminidase from human parainfluenza virus type III. *J Mol Biol.* **335**, 1343–1357 (2004).
34. Billaud, G. *et al.* Human parainfluenza virus type 4 infections: a report of 20 cases from 1998 to 2002. *J Clin Virol.* **34**, 48–51 (2005).
35. Fry, A. & Curns, A. Seasonal trends of human parainfluenza viral infections: United States, 1990-2004. *Clin Infect Dis.* **43**, 1016–1022 (2006).
36. Castaneda, S. J. & Wong, T. C. Leader sequence distinguishes between translatable and encapsidated measles virus RNAs. *J Virol.* **64**, 222–30 (1990).
37. Calain, P. & Roux, L. The rule of six, a basic feature for efficient replication of Sendai virus defective interfering RNA. *J Virol.* **67**, 4822–30 (1993).
38. Chang, A. & Dutch, R. E. Paramyxovirus fusion and entry: multiple paths to a common end. *Viruses.* **4**, 613–36 (2012).
39. Takimoto, T., Taylor, G., Connaris, H., Crennell, S. & Portner, A. Role of the hemagglutinin-neuraminidase protein in the mechanism of paramyxovirus-cell membrane fusion. *J Virol.* **76**, 13028–13033 (2002).
40. Gonzalez, Y. *et al.* Pulmonary enterovirus infections in stem cell transplant recipients. *Bone Marrow Transplant.* **23**, 511–513 (1999).
41. Munch, M., Nielsen, L. P., Handberg, K. J. & Jørgensen, P. H. Detection and subtyping (H5 and H7) of avian type A influenza virus by reverse transcription-PCR and PCR-ELISA. *Arch Virol.* **146**, 87–97 (2001).
42. Christensen, M. S., Nielsen, L. P. & Hasle, H. Few but severe viral infections in children with cancer: a prospective RT-PCR and PCR-based 12-month study. *Pediatr Blood Cancer.* **45**, 945–51 (2005).
43. Kapoor, A. *et al.* A highly prevalent and genetically diversified Picornaviridae genus in south Asian children. *Proc Natl Acad Sci U S A.* **105**, 20482–7 (2008).
44. Ning, Z., Cox, a. J. & Mullikin, J. C. SSAHA: a fast search method for large DNA databases. *Genome Res.* **11**, 1725–9 (2001).
45. Rasko, D. *et al.* Origins of the E. coli Strain causing an outbreak of hemolytic–uremic syndrome in Germany. *N Engl J Med.* **365**, 709–717 (2012).
46. Drummond, A. J. *et al.* Geneious v5.5, Available from http://www.geneious.com. Last accessed 31st August 2012 (2010).
47. Rozen, S. & Skaletsky, H. Primer3 on the WWW for General Users and for Biologists programmers. *Bioinformatics methods and protocols* (Misener, S. & Krawetz, S. A.) **132**, 365–386. (Humana Press Inc., 1999). Last accessed 31st August 2012.
48. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics.* **25**, 2078–9 (2009).
49. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. {MAFFT}: a novel method for rapid multiple sequence alignment based on fast {Fourier} transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).
50. Xia, X. & Xie, Z. DAMBE: Software package for data analysis in molecular biology and evolution. *J Hered.* 371–373 (2001).
51. Keane, T. M., Creevey, C. J., Pentony, M. M., Naughton, T. J. & Mclnerney, J. O. Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. *BMC Evol Biol.* **6**, 29 (2006).
52. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol.* **59**, 307–21 (2010).
53. Van de Peer, Y., Frickey, T., Taylor, J. & Meyer, A. Dealing with saturation at the amino acid level: a case study based on anciently duplicated zebrafish genes. *Gene.* **295**, 205–11 (2002).

54. Darriba, D., Taboada, G., Doallo, R. & Posada, D. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics.* **27**, 1164–1165 (2011).
55. Le, S. Q. & Gascuel, O. An improved general amino acid replacement matrix. *Mol Biol Evol.* **25**, 1307–20 (2008).

Acknowledgements

This project was made possible through funding by the Danish National Research Foundation, Lundbeck Foundation and The Kornerup Foundation. We also thank additional funding sources, The Danish Council for Independent Research (Technology and Production Sciences) and The Danish Council for Strategic Research (Program Commission on Strategic Growth Technologies).

Author contributions

D.E.A.P., T.M., C.A.W.B., A.J.H., Y.G., E.P., E.L.D., S.M., J.G. and L.P.N. conceived and designed the experiments. D.E.A.P., S.N.V., H.A.N., Y.G. and T.S. undertook the laboratory experiments. D.E.A.P., T.M., C.A.W.B., S.M., Y.G., A.S. and E.P. performed the data analysis. A.J.H., J.G., Y.G., A.S., E.P., E.L.D., L.P.N. contributed reagents, materials, and/or analysis tools. D.E.A.P. wrote the manuscript, with critical input from all other authors.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Alquezar-Planas, D.E. *et al*. Discovery of a divergent HPIV4 from respiratory secretions using second and third generation metagenomic sequencing. *Sci. Rep.* **3**, 2468; DOI:10.1038/srep02468 (2013).