Etiology of fever in Ugandan children: identification of microbial pathogens using metagenomic next-generation sequencing and IDseq, a platform for unbiased metagenomic analysis

Akshaya Ramesh[1, 2, † *], Sara Nakielny[3 *], Jennifer Hsu[4], Mary Kyohere[5], Oswald Byaruhanga[5], Charles de Bourcy[6], Rebecca Egger[6], Boris Dimitrov[6], Yun-Fang Juan[6], Jonathan Sheu[6], James Wang[6], Katrina Kalantar[3], Charles Langelier[4], Theodore Ruel[7], Arthur Mpimbaza[8], Michael R. Wilson[1, 2], Philip J. Rosenthal[4], Joseph L. DeRisi[3, 6 †]

[*] - These authors contributed equally to this work

[†] - Corresponding authors

1 – Weill Institute for Neurosciences, University of California, San Francisco, CA, USA

2 – Department of Neurology, University of California, San Francisco, CA, USA

3 – Department of Biochemistry and Biophysics, University of California, San Francisco, CA, USA

4 – Division of Infectious Diseases, Department of Medicine, University of California, San Francisco, CA, USA

5 – Infectious Diseases Research Collaboration, Kampala, Uganda

6 – Chan Zuckerberg Biohub, San Francisco CA

7 – Division of Pediatric Infectious Diseases and Global Health, Department of Pediatrics, University of California, San Francisco, CA, USA

8 – Child Health and Development Centre, Makerere University, Kampala, Uganda

e-mail address for authors:

Akshaya Ramesh: Akshaya.Ramesh@ucsf.edu

Sara Nakielny: ucsfsara@gmail.com

Jennifer Hsu: Jennifer.Hsu@kp.org

Mary Kyohere: pk_marx@yahoo.com

Oswald Byaruhanga: byaruos@gmail.com

Charles de Bourcy: cdebourcy@chanzuckerberg.com

Rebecca Egger: regger@chanzuckerberg.com

Boris Dimitrov: bdimitrov@chanzuckerberg.com

Yun-Fang Juan: yunfang@chanzuckerberg.com

Jonathan Sheu: jsheu@chanzuckerberg.com

James Wang: jwang@chanzuckerberg.com

Katrina Kalantar: kkalantar@berkeley.edu

Charles Langelier: Chaz.Langelier@ucsf.edu

Theodore Ruel: Theodore.Ruel@ucsf.edu

Arthur Mpimbaza: arthurwakg@yahoo.com

Michael R. Wilson: Michael.Wilson@ucsf.edu

Philip J. Rosenthal: Philip.Rosenthal@ucsf.edu

Joseph L. DeRisi: joe@derisilab.ucsf.edu

Abstract

Background: Febrile illness is a major burden in African children, and non-malarial causes of fever are uncertain. We built and employed IDseq, a cloud-based, open access, bioinformatics platform and service to identify microbes from metagenomic next-generation sequencing of tissue samples. In this pilot study, we evaluated blood, nasopharyngeal, and stool specimens from 94 children (aged 2-54 months) with febrile illness admitted to Tororo District Hospital, Uganda.

Results: The most common pathogens identified were *Plasmodium falciparum* (51.1% of samples) and parvovirus B19 (4.4%) from blood; human rhinoviruses A and C (40%), respiratory syncytial virus (10%), and human herpesvirus 5 (10%) from nasopharyngeal swabs; and rotavirus A (50% of those with diarrhea) from stool. Among other potential pathogens, we identified one novel orthobunyavirus, tentatively named Nyangole virus, from the blood of a child diagnosed with malaria and pneumonia, and Bwamba orthobunyavirus in the nasopharynx of a child with rash and sepsis. We also identified two novel human rhinovirus C species.

Conclusions: This exploratory pilot study demonstrates the utility of mNGS and the IDseq platform for defining the molecular landscape of febrile infectious diseases in resource limited areas. These methods, supported by a robust data analysis and sharing platform, offer a new tool for the surveillance, diagnosis, and ultimately treatment and prevention of infectious diseases.

Keywords: Metagenomic Next-Generation Sequencing, Febrile illness, surveillance, pathogen discovery, IDseq, cloud computing

Introduction

The evaluation of children with fever is challenging, particularly in the developing world. A febrile child in sub-Saharan Africa may have a mild self-resolving viral infection or may be suffering bacterial sepsis or malaria—major causes of disability and death [1], [2]. In the past, febrile illness in children under five years of age in most of Africa was treated empirically as malaria due to the limited availability of diagnostics and the risk of untreated malaria progressing to life-threatening illness. This strategy changed with new guidelines from the World Health Organization (WHO) in 2010, which recommend limiting malaria therapy to those with a confirmed diagnosis [3]. However, standard recommendations for management of febrile children who do not have malaria are lacking. Increased knowledge about the prevalence of non-malarial pathogens associated with fever is needed to inform management strategies for febrile children who do not have malaria [2].

Advances in genome sequencing hold promise for addressing global infectious disease challenges by enabling unbiased detection of microbial pathogens without requirement for the extensive infrastructure of modern microbiology laboratories [4], [5]. Although sequence-based diagnostics have not yet replaced most traditional microbiological assays, this situation is rapidly changing, as sequence-based strategies are incorporated for clinical care [6]–[10], and costs have declined dramatically over the past decade [11]. A significant roadblock toward implementation of sequence-based diagnostics is the extensive computational requirements and bioinformatics expertise required. In fact, as sequencing costs decline, computational expenses may proportionally increase due to inevitable expansion of existing genomic databases.

To address these challenges, we developed IDseq, a cloud-based open-source bioinformatics platform and service for detection of microbial pathogens from metagenomic next-generation sequencing (mNGS) data. IDseq requires minimal computational hardware and is designed to enhance accessibility and build informatics capacity in resource limited regions. Here, we leveraged IDseq and mNGS data to perform an exploratory proof-of-concept molecular survey of febrile children in rural Uganda to characterize pathogens associated with fever, including both well-recognized and novel causes of illness.

Analyses/Results

Clinical presentations of children providing samples for analysis

From October to December, 2013, 94 children admitted to Tororo District Hospital were enrolled (Table 1). Their mean age was 16.4 (IQR: 8.0-12.0) months, and 66 (70.2%) were female. Chief symptoms reported in addition to fever were cough (88.3%), vomiting (56.4%), diarrhea (47.9%), and convulsions (27.7%). Top admitting diagnoses were respiratory tract infection (57.4%), gastroenteritis/diarrhea (29.8%), and septicemia (11.7%) (Data file 1a).

Serum and nasopharyngeal (NP) swab samples were collected from 90 children each; for four children, only one of the two sample types was successfully collected. Although 45 (47.9%) of the children had a presenting symptom of diarrhea, stool samples were available for only 10 due to logistical constraints. Blood smears identified *P. falciparum* in 12 of the 90 samples that underwent mNGS analysis.

Metagenomic sequencing findings

mNGS was performed on 90 serum, 90 NP swab, and 10 stool samples and analyses was done using the IDseq pipeline (see Methods section, Figure 1); detailed findings are reported in Data file S1b. For each sample type, RNA was extracted, libraries were prepared, samples were sequenced, and reads were analyzed using the IDseq Portal. A mean of 11.5 million (IQR 6.4 – 15.2 million) paired-end reads were obtained per sample; sequencing statistics are in Data file Table S2. For one batch of serum samples, only a single read, rather than paired-end reads, was produced.

mNGS of serum

At least one microbial species was detected in 60 (66.7%) serum samples; more than one microbe was detected in 11 (12.2%) samples (Figure 2A). The most commonly identified microbes were *Plasmodium falciparum* (46, 51.1%) and parvovirus B19 (4, 4.4%). *P. falciparum* was detected in 10/12 samples from patients reported as smear-positive for malaria parasites. mNGS detected *Plasmodium spp.* in 37 additional samples that were smear negative (36 *P. falciparum* , 1 *P. malariae*). Viruses detected from serum included human immunodeficiency 1 virus (HIV-1), hepatitis A virus, rotavirus A, human herpesvirus (HHV) type 6, HHV type 4, HHV type 7, human rhinovirus (HRV)-C, HRV-A, enteroviruses (enterovirus A71, Coxsackievirus B2 and echovirus E30), human parechovirus 2, hepatitis B virus, a novel orthobunyavirus (described in greater detail below), human cardiovirus (Saffold virus), mamastrovirus 1 and Norwalk virus (Figure 2A).

Multiple viruses were detected from serum in patients with *Plasmodium* infections (10 of 46 (21.7%) samples; Table S1a). Three of the four identified parvovirus B19 infections were

associated with *P. falciparum*. Additionally, GB virus C and torque teno virus (TTV), which are of unknown clinical significance [12], [13], were identified in the serum of 25 (27.8%) and 37 (41.1%) children, respectively.

mNGS of NP swabs

A total of 90 NP swabs was collected and processed; 52 (57.7%) of these were from patients with admission diagnoses of pneumonia, respiratory tract infection, or bronchiolitis (Table 1). Chest imaging was not available to further assess these diagnoses. 73 NP samples (81.1%) contained at least 1 viral species (Figure 2B). HRV-A and HRV-C were the most prevalent, followed by respiratory syncytial virus (RSV), cytomegalovirus (HHV-5), influenza B, and coronavirus OC43. Other respiratory viruses identified included influenza A, HRV-B, adenovirus B, 3 human parainfluenza virus types, metapneumovirus, coronavirus NL63, avian coronavirus, coxsackievirus A2, coxsackievirus B2, polyomaviruses (KI), HHV-6 and HHV- 7. Other viruses identified that are not typically considered respiratory pathogens included hepatitis A virus, hepatitis B virus, parvovirus B19, mamastrovirus 1, Bwamba orthobunyavirus, betapapillomavirus 1, and rotavirus. Additionally, TTV was found in 49 (54.4%) NP swab samples, including one sample with both gemykrogvirus and TTV. For 26 (28.8%) patients, mNGS identified respiratory viral co-infections, most commonly with HRV-C (n=11) and HRV-A (n=5) (Table S1b). The same microbial species was identified in the NP swab and serum samples in 6 patients, one each with HRV-A, HRV-C, hepatitis A virus, hepatitis B virus, rotavirus A, and parvovirus B19.

Bacteria identified in NP samples included four dominant genera, which together comprised 79% of all microbial reads—*Moraxella* (39.4%), *Haemophilus* (16.7%), *Streptococcus* (16.2%), and *Corynebacterium* (6.6%). Given that diversity loss in the microbial flora in lower respiratory tract samples correlates with pneumonia [14], [15], we compared the Simpsons diversity Index (SDI) in patients with and without clinical diagnoses of respiratory tract infection. We found no significant difference between patients with (mean SDI = 0.51, IQR 0.37 - 0.65) or without (mean SDI = 0.51, IQR = 0.42 - 0.65; p = 0.86) diagnoses of respiratory infection (Figure S1). This finding is consistent with a growing body of work demonstrating that microbial composition of the nasopharynx may not correlate well with that of the lower respiratory tract in patients with pneumonia [16]–[18].

mNGS of stool

Among the 10 stool samples collected and sequenced, pathogens were detected in 9/10 samples. The three most common pathogens identified were rotavirus A (50%), *Cryptosporidium* (40%), and human parechovirus (40%). Seven children had co-infections: two rotavirus A and human parechovirus, and one each rotavirus A and *Cryptosporidium*, rotavirus A and enterovirus, *Cryptosporidium* and human parechovirus, *Cryptosporidium* and Norwalk virus, and *Giardia* and human parechovirus. Five samples also had *Blastocystis hominis*, a microbe of uncertain pathogenicity.

Orthobunyaviruses

Serum from one patient admitted with a clinical diagnosis of malaria and pneumonia contained a novel orthobunyavirus in addition to *P. falciparum*. Assembly of a draft genome and

comparison with existing orthobunyavirus genomes indicated that this draft sequence includes 97.5%, 100% and 91% of the L, M, and S coding regions, respectively (Fig 3). Average read coverage across the segments was 86-fold. Phylogenetic comparison showed that the novel virus was significantly divergent from known orthobunyaviruses, sharing 44.9-55.1% amino acid identity with the closest known relatives, Calchaqui virus, Kaeng Khoi virus, and Anopheles A virus (Figure 3, Figures S3a-c). Of note, despite the divergence of this virus, 47.9% of the reads that belong to this new genome were detectable using default parameters in the IDseq pipeline, allowing for facile subsequent assembly. The virus was isolated from a patient from Nyangole village, Tororo District—hence, we propose the name "Nyangole virus", consistent with nomenclature guidelines for the family *Bunyaviridae*.

In addition, a second orthobunyavirus, Bwamba virus, was identified in the NP swab sample from a patient admitted with rash, sepsis, and diarrhea. Insufficient sample and sequencing reads precluded genome assembly of this virus.

Genomic characterization of human rhinoviruses and influenza B in NP swabs

Human rhinoviruses

Within the species rhinovirus, we assembled *de novo* a total of 13 HRV-C (mean coverage: 39-fold) and 13 HRV-A (mean coverage: 268-fold) genomes (> 500 bp (Figure 4)). Of these, 10 HRV-A and 9 HRV-C genomes had complete coverage of the VP1 region, which is used to define enterovirus types [19]. HRV types are defined by divergence of >73% in the VP1 gene. As such, we found three HRV-A and eight HRV-C types in this cohort. One individual harbored two distinct HRV-A types (genome pairwise identity=75.3%, VP1 pairwise

identity=67.1%). Additionally, we assembled two novel HRV-C isolates from two patients admitted with gastroenteritis (patient ID: EOFI-014), and pneumonia, malaria and diarrhea (patient ID: EOFI-133), that shared 70.1% and 70.7% nucleotide sequence identity at VP1 compared to the closest known HRV-C (Accession JQ245968 and KF688606, respectively). The Picornavirus Working Group has established that novel HRV-Cs should exhibit at least 13% nucleotide sequence divergence in the VP1 gene [20], qualifying these two isolates as novel.

Influenza B virus

We assembled influenza B genome segments (>500bp, mean-coverage: 41-fold) from six of seven samples containing influenza B virus (one sample had insufficient sequencing reads). The viruses assembled were >99% similar to each other and >99% identical to the B/Massachusetts/02/2012-like virus included in the vaccine recommended by WHO for the 2013-2014 northern hemisphere and 2014 southern hemisphere influenza seasons (Accession numbers: NC_002204 to NC_002211).

Antimicrobial resistance profiling in the nasopharynx and stool

We employed Short Read Sequencing Typing (SRST2) to survey the resistance gene landscape of our metagenomic dataset [70]. In total, 86% of patients were found to harbor molecular evidence of resistant organisms in their nasopharynx and 50% in their stool. For both stool and NP swabs, genes conferring resistance to beta lactams were the most abundant, and included *ampC* (n=2), which confers resistance to first through third generation cephalosporins, and *IMP-1* (n=1) which confers broad spectrum resistance that includes carbapenems. In

addition, genetic signatures of resistance to other antibacterial agents including aminoglycosides, macrolides/lincosamides, phenicols, sulfas, and tetracyclines were identified (Figure S2).

Discussion

The clinical management of children with fever is challenging in Africa, where clinicians often have access only to malaria diagnostics [21]. A better understanding of the microbial agents causing fever in African children is needed to inform the development of better diagnostic algorithms and therapeutic guidelines. To address this unmet need, we developed and deployed IDseq in combination with unbiased mNGS to characterize the etiology of fever in Ugandan children admitted to a rural district hospital. We identified a wide range of pathogens in these children.

Other studies evaluating causes of febrile illness in African children have focused on a limited number of pathogens [22]–[25]. In a study of febrile children in Tanzania utilizing rapid diagnostic tests, serologic tests, culture, and molecular tests, viruses accounted for 51% of lower respiratory infections, 78% of systemic infections, and 100% of nasopharyngeal infections. Additionally, 9% of the children had malaria and 4.2% bacteremia [26]. In febrile children in Kenya, reported pathogens were spotted fever group *Rickettsia* (22.4%), influenza (22.4%), adenovirus (10.5%,) parainfluenza virus 1-3 (10.1%), Q fever (8.9%), RSV (5.3%), malaria (5.2%), scrub typhus (3.6%), human metapneumovirus (3.2%), group A *Streptococcus* (2.3)% and typhus group Rickettsiae (1.0%) [27], [28]. Another study reported bacteremia in 19.1% of children admitted to a referral hospital in Uganda [29]. Additionally, in patients (across all age

groups) with severe febrile illness, bacteremia was detected in 10.1% in North Africa, 10.4% in East Africa, and 12.4% in West Africa [30].

Traditional pathogen detection methods, including culture, serology, and pathogen directed molecular methods, are logistically challenging in resource limited settings due to the need for extensive microbiology laboratory infrastructure. Unbiased sequencing approaches are designed to identify all potential pathogens, but have been limited by high cost. The costs of deep sequencing are rapidly decreasing, but the analysis of mNGS data necessarily incurs an increasingly large cost, as the available genomic databases to be searched continue to expand. Other major challenges for mNGS approaches include the need for better control datasets, in particular to enable discrimination of contaminants and commensal organisms from true pathogens. The IDseq platform was based on our prior experience with in-house pipelines for pathogen detection [31]–[33], and it aims to address existing computational and bioinformatics barriers by providing facile cloud-based mNGS analysis without the requirement for significant on-premise computational infrastructure.

Using mNGS and IDseq, the most common pathogen identified in the blood of febrile Ugandan children was *P. falciparum*, an expected result considering the high incidence of malaria in Tororo District at the time of this study [34]. Some discrepancies were seen compared to blood smear readings, with false positive smears probably due to errors in slide reading, a common problem in African clinics [35], and false negative smears due to the expected greater sensitivity of mNGS for identification of *P. falciparum*. In children with only sub-microscopic parasitemia, it is uncertain whether fevers can be ascribed to malaria, and in fact many children

had both *P. falciparum* and additional pathogens identified. Interestingly, three of the four cases of parvovirus B19 were found in association with *P. falciparum*; this co-infection has been associated with severe anemia with life-threatening consequences [36]–[38].

HRV was the most commonly identified virus in NP swab samples, consistent with findings previously reported in sub-Saharan Africa and developed countries [39]–[43]. HRV-C was most frequently encountered (54.1%), followed by HRV-A (43.2%) and HRV-B (2.7%), similar to the distribution of HRVs previously reported in Kenya [39]. We identified two novel HRV-C species; these viruses were about 70% identical to the most closely related previously described HRV-C species [20]. Overall, we detected at least three HRV-A and eight HRV-C types co-circulating in Tororo District. Of note, during the same collection period, a lethal HRV-C outbreak was reported in chimpanzees in Kibale National Park, in western Uganda [44]. The HRV-C reported in western Uganda was modestly related to an isolate observed in our study (74% nucleotide identity; 81% amino acid identity) (Figure 4) [44]. Our results confirm that a wide spectrum of HRVs infects Ugandan children. In addition to HRV, we detected a spectrum of other known respiratory viruses, including RSV, human parainfluenza viruses, human coronaviruses, and adenovirus.

Diarrheal disease is one of the leading causes of death in children in Africa [45]. Approximately 48% of febrile children in our study presented with diarrhea, but due to logistical constraints stool specimens were available for only 10 cases. Rotavirus A, the leading cause of pediatric diarrhea worldwide [46], was the most commonly identified pathogen in this cohort. Rotavirus vaccination, known to be highly effective, is yet to be implemented in Uganda, but the

need is clear [46]. In addition to rotavirus A, we detected *Cryptosporidium*, norovirus, *Giardia*, *B. hominis* and several enteroviruses in stool specimens. Enteroviruses, HRV-C, and mamastrovirus were also identified in the serum of three children with clinical diagnoses of gastroenteritis or diarrhea.

Unbiased inspection of microbial sequences from sera revealed a novel member of the orthobunyavirus genus, tentatively named Nyangole virus, which was identified as a co-infection with *P. falciparum* in a child with clinical diagnoses of malaria and pneumonia. The virus was surprisingly divergent from known viruses, with an average amino acid similarity of 51.6% to its nearest known relatives including Calchaqui, Anopheles A and Kaeng Khoi viruses. Mosquitoes have been proposed as a vector for Calchaqui and Anopheles A viruses; Kaeng Khoi virus has been isolated from bedbugs [47]–[49]. Antibodies to these viruses have been detected in human sera, but their role as human pathogens is uncertain [47]–[50]. While the depth of coverage of Nyangole virus sequence in our patient suggests significant viremia, and other orthobunyaviruses are responsible for severe human illnesses (e.g., California encephalitis virus, La Crosse virus, Jamestown Canyon virus, and Cache Valley virus) [51], it is unknown whether the identified virus was responsible for the presenting symptoms.

NP swab analysis identified another Orthobunyavirus, Bwamba virus, in a child admitted with rash, sepsis and diarrhea. This virus has previously been described to cause fever in Uganda [41]. Our identification of two orthobunyaviruses, including one novel virus, in a small sample of febrile Ugandan children suggests that the landscape of previously unidentified viruses that infect African children and potentially cause febrile illness, is significantly under explored.

Molecular evidence of resistance to every available antibiotic except vancomycin in the WHO Essential Medicines List was identified in this study [52]. Three children had genotypic evidence of ESBL producing organisms predicted to be resistant to ceftriaxone, a frontline antibiotic for severe infections in this region [53]. In addition, one of these three children carried IMP-1, which also confers resistance to carbapenems, a class of antibiotics reserved for the most resistant infections.

Our exploratory pilot study had important limitations. First, our samples were not collected randomly, but rather were a convenience sample due to the logistical constraints of our small clinical study; as such the results should not be seen as broadly representative of pathogens infecting Ugandan children. In particular, the lack of identification of bacteremia in study subjects may have been due to a relative paucity of severe illness, compared to that in other studies, in our cohort of admitted children. Second, clinical evaluation of children followed the standards of a rural African hospital, so diagnostic evaluation was limited to physical examination and malaria blood smears. Much more may be learned by linking rigorous clinical evaluation with mNGS results, and thereby more comprehensively assessing associations between clinical syndromes and specific pathogens.

Despite these limitations, our study demonstrates the utility of the mNGS/IDseq platform, and provides an important cross sectional snapshot of causes of fever in African children. Combining unbiased mNGS and the IDseq platform enabled the identification of likely known and novel causes of pediatric illness, and makes available a powerful new open access tool for the characterization of infectious diseases in resource limited settings

Potential implications

This study provides a snapshot of pediatric fever in Tororo District, Uganda. mNGS permits universal pathogen detection using a single assay, and thus avoids the need for multiple independent, and often costly tests to determine disease etiology. Despite the promise and decreasing costs of this technology, the extensive computational and bioinformatic infrastructure needed to perform analysis of sequencing data remains a major barrier to implementation of mNGS in the developing world. Here we address the need for bioinformatic democratization and computational capacity building with IDseq, an open access platform to bring infectious disease surveillance to regions where it is needed most. As progress is made toward elimination of malaria in sub-Saharan Africa, it will be increasingly important to understand the landscape of pathogens that account for the remaining burden of morbidity and mortality. The use of mNGS can contribute importantly to this understanding, offering unbiased identification of infecting pathogens.

Materials and Methods

Enrollment of study subjects:

We studied children admitted to Tororo District Hospital, Tororo, Uganda, with febrile illnesses. Potential subjects were identified by clinic staff, who notified study personnel, who subsequently evaluated the children for study eligibility. Inclusion criteria were: 1) age 2-60 months; 2) admission to Tororo District Hospital for acute illness; 3) documentation of axillary temperature >38.0°C on admission or within 24 hours of admission; and 4) provision of informed consent from the parent or guardian for study procedures.

Study specimens:

NP swabs and blood were collected from each enrolled subject within 24 hours of hospital admission. Approximately 5 ml of blood was collected by phlebotomy, the sample was centrifuged at room temperature, and serum was then stored at -80°C. NP swab samples collected with FLOQSwabs™ swabs (COPAN) were placed into cryovials with Trizol (Invitrogen), and stored at -80°C within ~5 min of collection. For subjects with acute diarrhea (≥ 3 loose or watery stools in 24 hours), stool was collected into clean plastic containers and stored at -80°C within ~5 min of collection. Samples were stored at -80°C until shipment on dry ice to UCSF for sequencing.

Clinical data:

Clinical information was obtained from interviews with parents or guardians, with specific data entered onto a standardized case record form that included admission diagnosis and physical examination as well as malaria blood smear results. For malaria diagnosis, thick blood smears were Giemsa stained and evaluated by Tororo District Hospital laboratory personnel following routine standard-of-care practices. No efforts were made to improve on routine practice, so malaria smear readings represent routine standard-of-care rather than optimal quality controlled reads.

Metagenomic next-generation sequencing (mNGS):

After shipment to University of California, San Francisco, RNA was extracted from clinical samples as well as positive (HeLa cells) and negative (water) controls, and unbiased

cDNA libraries were generated using previously described methods [31], [54], [55]. Barcoded samples were pooled, size selected (Blue Pippin), and run on an Illumina HiSeq2500 to obtain 135 base pair (bp) paired-end reads

Bioinformatic analysis and pathogen identification:

Microbial pathogens were identified from raw sequencing reads using the IDseq Portal (https://idseq.net), a novel cloud-based, open-source bioinformatics platform designed for detection of microbes from metagenomic data (Figure 1). IDseq scripts and user instructions are available at https://github.com/chanzuckerberg/idseq-dag and the graphical user interface web application for sample upload is available at https://github.com/chanzuckerberg/idseq-web. IDseq is conceptually based on previously implemented platforms [31]–[33], but is optimized for scalable Amazon Web Services (AWS) cloud deployment. Bioinformatics data processing jobs are carried out on demand as Docker containers using AWS Batch. Alignments to the National Center for Biotechnology Information (NCBI) database are executed on dedicated auto scaling groups (ASG) of Amazon Elastic Compute Cloud (EC2) instances, with the number of server instances varied with job load. Fast downloads of the NCBI database from the Amazon Simple Storage Service to each new server instance are enabled by the open-source tool s3mi (https://github.com/chanzuckerberg/s3mi). Initial alignment and removal of reads derived from the human genome is performed using the Spliced Transcripts Alignment to a Reference (STAR) algorithm [56]. Low-quality reads, duplicates, and low-complexity reads are then removed using the Paired-Read Iterative Contig Extension (PRICE) computational package [57], the CD-HIT-DUP tool [58], and a filter based on the Lempel-Ziv-Welch (LZW) compression score, respectively. A second round of human read filtering is carried out using bowtie2 [59].

Remaining reads are queried against the most recent version of the NCBI nucleotide and non-redundant protein databases (updated monthly) using GSNAPL and RAPSearch2 [60], [61], respectively. Reads matching GenBank records in the superphylum Deuterostomia are removed, given the high likelihood that such residual reads are of human origin. The relative abundance of microbial taxa is calculated based on reads per million (rpM) mapped at the genus level. To distinguish potential pathogens from ubiquitous environmental agents including commensal flora, a Z-score is calculated for the value observed for each genus relative to a background of healthy and no-template control samples [32]. An overview of this pipeline is represented in Figure 1.

IDseq can process 150 samples at a given time. As of this writing, the current version of the IDseq pipeline (IDseqv1.8) processes fastq files with approximately 70 million reads, typically containing >99% host sequence, in 34 minutes. Run times may vary depending on demand, percentage of non-host sequence, and autoscaling parameters.

For this study we report species greater than 0 rpM and Z-scores detected in the serum, stool, and NP samples. Consistent with previous studies, low levels of “index bleed through” or “barcode hopping” (assignment of sequencing reads to the wrong barcode/index) was observed within the non-templated control samples [62]. To prevent mis-assignment, when a microbe found in more than one sample, it was reported only when present at levels at least four times the level of mis-assigned reads observed in the control samples. Given the extremely high levels of rotavirus found in stool samples, these samples were run in duplicate, and only microbes identified in both samples and present at levels at least four times the number of reads mis-

assigned in the control samples were reported. If the reads identified for a given pathogen were not species-specific, we reported the corresponding genus. For NP and stool samples, because the nasopharynx and intestines are normally colonized with commensal bacteria [63]–[66], only non-bacterial species were reported.

Genome assembly, annotation and phylogenetic analysis:

To more comprehensively characterize the genomes of identified microbes PRICE [57] and St. Petersburg genome assembler (SPAdes) [67] were used to *de novo* assemble short read sequences into larger contiguous sequences (contigs). Assembled contigs were queried against the NCBI nt database using BLAST to identify the closest related microbes. GenBank annotation files from genome sequence records corresponding to the highest scoring alignments were used to identify potential features within the *de novo* assembled genomes. Geneious v10.3.2 was used to annotate newly assembled genomes. Reference genomes for multiple sequence alignments and phylogenetic analyses were downloaded from NCBI. Multiple sequence alignments were generated using ClustalW in MEGA v6.0 and the Geneious aligner in Geneious v10.3.2. Neighbor-joining phylogenetic trees were generated using Geneious v10.3.2 and further assessed using FigTree v1.4.3. Annotation of protein domains in the novel orthobunyaviruses was performed using the InterPro webserver [68] as well as direct alignment against previously known orthobunyaviruses. The TOPCONS webserver [69] was used for the identification of transmembrane regions and signal peptides, and the NetNglyc 1.0 Server (http://www.cbs.dtu.dk/services/NetNGlyc/) for the identification of glycosylation sites.

Evaluation of NP microbiome diversity:

We applied SDI to evaluate alpha diversity of microbes identified in NP samples. For this analysis patients were stratified into two categories based on clinical assignment: respiratory infections (admitting diagnosis of pneumonia, respiratory tract infection, or bronchiolitis; n=52) and all other syndromes (n=39); cases with unknown admitting diagnosis were excluded. SDI was calculated in R using the Veganv2.4.4 package on genus-level reads per million values for all microbes, including bacteria. A Wilcox Rank Sum test was used to evaluate differences in SDI between patients in the two categories.

Antimicrobial resistance gene identification in nasopharynx and stool:

The SRST2 computational package was used to identify antimicrobial resistance genes using the Argannot2 database as previously described [70]. We defined extended spectrum β-lactamase (ESBL) as a β-lactamase conferring resistance to the penicillins, first-, second-, and third-generation cephalosporins including the gene classes proposed by Giske et al [71].

Availability of data and code

All raw data have been deposited under Bioproject ID: PRJNA483304. Assembled genomes can be accessed in GenBank: Accession numbers: MH685676-MH685701, MH685703- MH685719, MH684286-MH684293, MH684298-MH684334. All the raw data, intermediate data and IDseq reports can also be accessed at https://idseq.net [Project ID: Uganda - all - 2]. All IDseq scripts and user instructions are available at https://github.com/chanzuckerberg/idseq-dag and the graphical user interface web application for sample upload is available at https://github.com/chanzuckerberg/idseq-web. While IDseq is

under continuous development and improvement, full version control of processing runs and reference databases are standard features.

Declarations:

List of abbreviations

AWS Amazon Web Services
ASG Auto Scaling Groups
BLAST Basic Local Alignment Search Tool
EC2 Elastic Compute Cloud
ESBL Extended spectrum β-lactamase
HHV human herpesvirus
HIV-1 human immunodeficiency 1 virus
HRV Human Rhinovirus
IQR Inter-quartile range
L coding region Large segment coding region (RNA-dependent RNA polymerase)
LZW Lempel-Ziv-Welch
M coding region Medium segment coding region (glycoprotein)
MEGA Molecular Evolutionary Genetics Analysis
mNGS metagenomic next-generation sequencing
NCBI National Center for Biotechnology Information
NP swab Nasopharyngeal swab
nr database non-redundant database
nt database nucleotide database
PRICE Paired-Read Iterative Contig Extension
rpM reads per million
RSV respiratory syncytial virus
S coding region Small segment coding region (nucleocapsid)
SDI Simpsons diversity Index
SPAdes St Petersburg genome assembler
SRST2 Short Read Sequencing Typing
STAR Spliced Transcripts Alignment to a Reference
TTV torque teno virus
VP1 Capsid protein VP1
WHO World Health Organization

Ethics approval and consent to participate

The study protocol was approved by the Uganda National Council of Science and Technology and the Institutional Review Boards of the School of Medicine, Makerere University-College of Health Sciences and the University of California, San Francisco.

Competing interests

None

Funding

JLD is supported by the Chan Zuckerberg Biohub. CB, BD, YJ, JS, RE and JW are funded by the Chan Zuckerberg Initiative. SN was supported by Howard Hughes Medical Institute. PJR is funded by the National Institutes of Health, and work on this project by JH, MK, OB, and PJR was funded by the Doris Duke Charitable Foundation. MRW is supported by the Rachleff Foundation, NINDS K08NS096117 and the University of California, San Francisco Center for Next-Gen Precision Diagnostics, which is supported by the Sandler Foundation and the William K. Bowes, Jr. Foundation. AR is supported by the Rachleff Foundation. CL is supported by NHLBI K23HL138461-01A1 , Nina Ireland Foundation and Marcus Program in Precision Medicine. KK is supported by the UC Berkeley UC San Francisco Joint Program in Bioengineering.

Authors contributions

SN, JH, MK, OB, TR, AM, PR and JLD contributed to experimental design, data acquisition and sample processing. SN, JH, MK, OB, PR, TR, and AM, contributed to patient recruitment and clinical testing. CB, BD, YJ, JS, RE, and JW developed the IDseq pipeline. AR, KK, CL, SN,

PR, MRW, and JLD contributed to data analysis. AR, MRW, PR, and JLD contributed to manuscript writing.

Acknowledgements

We thank the children and families in Tororo District, Uganda, who participated in this study. We thank Eric Chow, Jessica Lund, and the UCSF Center for Advanced Technology for assistance with sequencing. We recognize Amy Kistler. for intellectual discussions and comments on the paper. We thank Mark Stenglein for his helpful discussions and advice.

Figures:

Figure 1: Schematic representation of IDseq pipeline

Figure 2: Microbes found in (A) serum and (B) nasopharyngeal (NP) swab samples. Note that bacterial species were not considered for NP samples.

Figure 3: Characterization of the novel orthobunyavirus identified in a febrile child. (A) Schematic representation of the large (L) or RNA dependent RNA polymerase, medium (M) or polyprotein of Gn, NSm and Gc proteins and small (S) segment encoding the nucleocapsid (N) protein of Nyangole virus and percentage identity with the most closely related virus. Phylogenetic tree of all complete orthobunyavirus genome sequences along with Nyangole virus are represented in (B) for the RNA dependent RNA polymerase and (C) for the glycoprotein.

Figure 4: Phylogenetic tree of all complete HRV genomes from NCBI and HRV genomes assembled in this study (Purple – Rhinovirus A, Orange – Rhinovirus C)

Supplemental figures:

Figure S1: Simpsons diversity index (SDI) for samples with pneumonia versus other etiologies. Each triangle represents one sample.

Figure S2: Antimicrobial resistant genes identified in (A) nasopharynx and (B) stool samples

Figure S3: Complete phylogenetic tree of (A) Large (L) or RNA dependent RNA polymerase, (B) Medium (M) or polyprotein of Gn, NSm and Gc proteins and (C) Small (S) segment encoding Nucleocapsid segments

Table 1: Overview of the patients enrolled in the study

Clinical category	Age (mean, months)	Gender	
		Male	Female
Respiratory illness (54)	14.0	25	28
Diarrhea/gastroenteritis (28)	12.1	9	19
Malaria (11)	15.5	3	7
Sepsis (11)	19.9	5	6
Malnutrition (5)	18.4	2	3
Other (15)	21.1	11	3

The other category includes the following admission criteria: unknown (n=10), UTI (n=1), meningitis (n=2), hepatitis (n=1), fever (n=1). Gender information were missing for one child in the following categories: Respiratory illness, Malaria and Other. Age information were missing for one child in the following categories: Respiratory illness and Other.

Table S1: (A) Co-infection table for *P. falciparum*

Microbial co-infections with *P.falcipatum*	**Number of cases**
Parvovirus B19	2
Human herpesvirus 4	1
Human immunodeficiency virus 1	1
Norwalk virus	1
Orthobunyavirus	1
HRV-A	1
HRV-C	1
Rotavirus A	1
Parvovirus B19, HRV-C and human parechovirus 2	1

Table S1: B) Co-infection table for HRV

microbial co-infections with HRV	number of cases
Human coronavirus OC43	2
Human parainfluenza virus 1	2
Rotavirus A	2
Hepatitis A	1
HHV type 5	1
HHV type 6	1
Human parainfluenza virus 4	1
RSV	1
KI polyomavirus	1
RSV and Mamastrovirus 1	1
RSV and human parainfluenza virus 1	1
HHV type 5 HHV type 7, human coronavirus OC43	1
Hepatitis B, human coronavirus NL63 and influenza A	1

Data files:

Data file 1: (A) Clinical signs and symptoms of patients enrolled in the study. (B) mNGS findings in patients enrolled in the study

Data file 2: Total number of sequencing reads and unique non-human reads in all samples analyzed

References

1. Maze MJ, Bassat Q, Feasey NA, Mandomando I, Musicha P, Crump JA. The epidemiology of febrile illness in sub-Saharan Africa: implications for diagnosis and management. Clin Microbiol Infect. 2018. doi:10.1016/j.cmi.2018.02.011.
2. Prasad N, Sharples KJ, Murdoch DR, Crump JA. Community prevalence of fever and relationship with malaria among infants and children in low-resource areas. Am J Trop

Med Hyg. 2015;93(1):178-180. doi:10.4269/ajtmh.14-0646.

3. Guidelines for the treatment of malaria, 2nd edition. World Heal Organ. 2010:197p. doi:WHO Library Cataloguing-in-Publication Data.

4. Bibby K. Metagenomic identification of viral pathogens. Trends Biotechnol. 2013;31(5):275-279. doi:10.1016/j.tibtech.2013.01.016.

5. Yozwiak NL, Skewes-Cox P, Stenglein MD, Balmaseda A, Harris E, DeRisi JL. Virus identification in unknown tropical febrile illness cases using deep sequencing. PLoS Negl Trop Dis. 2012;6(2):e1485. doi:10.1371/journal.pntd.0001485.

6. Wilson MR, Naccache SN, Samayoa E, et al. Actionable Diagnosis of Neuroleptospirosis by Next-Generation Sequencing. N Engl J Med. 2014;370(25):2408-2417. doi:10.1056/NEJMoa1401268.

7. Wilson MR, Shanbhag NM, Reid MJ, et al. Diagnosing Balamuthia mandrillaris Encephalitis with Metagenomic Deep Sequencing. Ann Neurol. 2015;78(5):722-730. doi:10.1002/ana.24499.

8. Doan T, Wilson MR, Crawford ED, et al. Illuminating uveitis: metagenomic deep sequencing identifies common and rare pathogens. Genome Med. 2016;8:1-2. doi:10.1186/s13073-016-0344-6.

9. Quan PL, Wagner TA, Briese T, et al. Astrovirus encephalitis in boy with X-linked agammaglobulinemia. Emerg Infect Dis. 2010;16(6):918-925. doi:10.3201/eid1606.091536.

10. Palacios G, Druce J, Du L, et al. A New Arenavirus in a Cluster of Fatal Transplant-Associated Diseases Gustavo. N Engl J Med. 2008;358:991-998. doi:10.1056/NEJMoa1514204.

11. Muir P, Li S, Lou S, et al. The real cost of sequencing: Scaling computation to keep pace with data generation. Genome Biol. 2016;17(1):1-9. doi:10.1186/s13059-016-0917-0.

12. Naoumov N V. TT virus - highly prevalent, but still in search of a disease. J Hepatol. 2000;33(13):157-159.

13. Hafez MM, Shaarawy SM, Hassan AA, Salim RF, Abd El Salam FM, Ali AE. Prevalence of transfusion transmitted virus (TTV) genotypes among HCC patients in Qaluobia governorate. Virol J. 2007;4:1-6. doi:10.1186/1743-422X-4-135.

14. Abreu NA, Nagalingam NA, Song Y, et al. Sinus Microbiome Diversity Depletion and Corynebacterium tuberculostearicum Enrichment Mediates Rhinosinusitis. Sci Transl Med. 2012;4(151):151ra124-151ra124. doi:10.1126/scitranslmed.3003783.

15. Langelier C, Zinter M, Kalantar K, et al. Metagenomic Next-Generation Sequencing Detects Pulmonary Pathogens in Hematopoietic Cellular Transplant Patients with Acute Respiratory Illnesses. 2017.

16. Park DE, Baggett HC, Howie SRC, et al. Colonization density of the upper respiratory tract as a predictor of pneumonia - Haemophilus influenzae, Moraxella catarrhalis, Staphylococcus aureus, and Pneumocystis jirovecii. Clin Infect Dis. 2017;64(May):S328-S336. doi:10.1093/cid/cix104.

17. Loens K, Van Heirstraeten L, Malhotra-Kumar S, Goossens H, Ieven M. Optimal sampling sites and methods for detection of pathogen possibly causing community-acquired lower respiratory tract infections. J Clin Microbiol. 2009;47(1):21-31. doi:10.1128/JCM.02037-08.

18. Zar HJ, Barnett W, Stadler A, Gardner-Lubbe S, Myer L, Nicol MP. Aetiology of childhood pneumonia in a well vaccinated South African birth cohort: A nested case-

control study of the Drakenstein Child Health Study. Lancet Respir Med. 2016;4(6):463-472. doi:10.1016/S2213-2600(16)00096-5.

19. Lukashev AN, Vakulenko YA. Molecular evolution of types in non-polio enteroviruses. J Gen Virol. 2017;98(12):2968-2981. doi:10.1099/jgv.0.000966.

20. McIntyre CL, Knowles NJ, Simmonds P. Proposals for the classification of human rhinovirus species A, B and C into genotypically assigned types. J Gen Virol. 2013;94(PART8):1791-1806. doi:10.1099/vir.0.053686-0.

21. Inouye M, Dashnow H, Raven LA, et al. SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. Genome Med. 2014;6(11):1-16. doi:10.1186/s13073-014-0090-6.

22. Crump JA, Kirk MD. Estimating the Burden of Febrile Illnesses. PLoS Negl Trop Dis. 2015;9(12):1-7. doi:10.1371/journal.pntd.0004040.

23. Chipwaza B, Mugasa JP, Selemani M, et al. Dengue and Chikungunya Fever among Viral Diseases in Outpatient Febrile Children in Kilosa District Hospital, Tanzania. PLoS Negl Trop Dis. 2014;8(11). doi:10.1371/journal.pntd.0003335.

24. Jacob ST, Pavlinac PB, Nakiyingi L, et al. Mycobacterium tuberculosis Bacteremia in a Cohort of HIV-Infected Patients Hospitalized with Severe Sepsis in Uganda-High Frequency, Low Clinical Sand Derivation of a Clinical Prediction Score. PLoS One. 2013;8(8). doi:10.1371/journal.pone.0070305.

25. Crump JA, Morrissey AB, Nicholson WL, et al. Etiology of Severe Non-malaria Febrile Illness in Northern Tanzania: A Prospective Cohort Study. PLoS Negl Trop Dis. 2013;7(7). doi:10.1371/journal.pntd.0002324.

26. Chipwaza B, Mhamphi GG, Ngatunga SD, et al. Prevalence of Bacterial Febrile Illnesses

in Children in Kilosa District, Tanzania. PLoS Negl Trop Dis. 2015;9(5). doi:10.1371/journal.pntd.0003750.

27. D'Acremont V, Kilowoko M, Kyungu E, et al. Beyond Malaria — Causes of Fever in Outpatient Tanzanian Children. N Engl J Med. 2014;370(9):809-817. doi:10.1056/NEJMoa1214482.

28. O'Meara WP, Mott JA, Laktabai J, et al. Etiology of pediatric fever in Western Kenya: A case-control study of falciparum Malaria, Respiratory Viruses, and Streptococcal Pharyngitis. Am J Trop Med Hyg. 2015;92(5):1030-1037. doi:10.4269/ajtmh.14-0560.

29. Maina AN, Farris CM, Odhiambo A, et al. Q fever, scrub typhus, and rickettsial diseases in children, Kenya, 2011–2012. Emerg Infect Dis. 2016;22(5):883-886. doi:10.3201/eid2205.150953.

30. Kibuuka A, Byakika-Kibwika P, Achan J, et al. Bacteremia among febrile ugandan children treated with antimalarials despite a negative malaria test. Am J Trop Med Hyg. 2015;93(2):276-280. doi:10.4269/ajtmh.14-0494.

31. Prasad N, Murdoch DR, Reyburn H, Crump JA. Etiology of severe febrile illness in low- and middle-income countries: A systematic review. PLoS One. 2015;10(6):1-25. doi:10.1371/journal.pone.0127962.

32. Yozwiak NL, Skewes-Cox P, Stenglein MD, Balmaseda A, Harris E, DeRisi JL. Virus identification in unknown tropical febrile illness cases using deep sequencing. PLoS Negl Trop Dis. 2012;6(2). doi:10.1371/journal.pntd.0001485.

33. Wilson MR, O'Donovan BD, Gelfand JM, et al. Chronic Meningitis Investigated via Metagenomic Next-Generation Sequencing. JAMA Neurol. 2018;94158. doi:10.1001/jamaneurol.2018.0463.

34. Naccache SN, Federman S, Veeraraghavan N, et al. A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. Genome Res. 2014;24(7):1180-1192. doi:10.1101/gr.171934.113.

35. Oguttu DW, Matovu JKB, Okumu DC, et al. Rapid reduction of malaria following introduction of vector control interventions in Tororo District, Uganda: a descriptive study. Malar J. 2017;16(1):1-8. doi:10.1186/s12936-017-1871-3.

36. Mekonnen SK, Aseffa A, Medhin G, Berhe N, Velavan TP. Re-evaluation of microscopy confirmed Plasmodium falciparum and Plasmodium vivax malaria by nested PCR detection in southern Ethiopia. Malar J. 2014;13(48). doi:10.1186/1475-2875-13-48.

37. Agarwal R, Baid R, Datta R, Saha M, Sarkar N. Falciparum malaria and parvovirus B19 coinfection: A rare entity. Trop Parasitol. 2017;7(1):47-48.

38. Duedu KO, Sagoe KWC, Ayeh-Kumi PF, Affrim RB, Adiku T. The effects of co-infection with human parvovirus B19 and Plasmodium falciparum on type and degree of anaemia in Ghanaian children. Asian Pac J Trop Biomed. 2013;3(2):129-139. doi:10.1016/S2221-1691(13)60037-4.

39. Toan NL, Sy BT, Song LH, et al. Co-infection of human parvovirus B19 with Plasmodium falciparum contributes to malaria disease severity in Gabonese patients. BMC Infect Dis. 2013;13(1). doi:10.1186/1471-2334-13-375.

40. Onyango CO, Welch SR, Munywoki PK, et al. Molecular epidemiology of human rhinovirus infections in Kilifi, coastal Kenya. J Med Virol. 2012;84(5):823-831. doi:10.1002/jmv.23251.

41. O'Callaghan-Gordo C, Bassat Q, Morais L, et al. Etiology and epidemiology of viral pneumonia among hospitalized children in rural mozambique: A malaria endemic area

with high prevalence of human immunodeficiency virus. Pediatr Infect Dis J. 2011;30(1):39-44. doi:10.1097/INF.0b013e3181f232fe.

42. Mbayame Ndiaye Niang, Ousmane M. Diop, Fatoumata Diene Sarr, Deborah Goudiaby, Hubert Malou-Sompy, Kader Ndiaye, Astrid Vabret LB. Viral Etiology of Respiratory Infections in Children Under 5 Years Old Living in Tropical Rural Areas of Senegal: The EVIRA Project. J Med Virol. 2010;82:866-872. doi:10.1002/jmv.

43. Smuts HE, Workman LJ, Zar HJ. Human rhinovirus infection in young African children with acute wheezing. BMC Infect Dis. 2011;11(1):65. doi:10.1186/1471-2334-11-65.

44. Jain S, Self WH, Wunderink RG, et al. Community-Acquired Pneumonia Requiring Hospitalization among U.S. Adults. N Engl J Med. 2015;373(5):415-427. doi:10.1056/NEJMoa1500245.

45. Scully EJ, Basnet S, Wrangham RW, et al. Lethal Respiratory Disease Associated with Human Rhinovirus C in Wild Chimpanzees, Uganda, 2013. Emerg Infect Dis. 2018;24(2). doi:10.3201/eid2402.170778.

46. World Health Organization. Diarrhoeal Disease.; 2017. http://www.who.int/news-room/fact-sheets/detail/diarrhoeal-disease.

47. Mwenda JM, Burke RM, Shaba K, et al. Implementation of Rotavirus Surveillance and Vaccine Introduction - World Health Organization African Region, 2007-2016. MMWR Morb Mortal Wkly Rep. 2017;66(43):1192-1196. doi:10.15585/mmwr.mm6643a7.

48. Calisher CH, Monath TP, Sabattini MS, et al. A newly recognized vesiculovirus, Calchaqui virus, and subtypes of Melao and Maguari viruses from Argentina, with serologic evidence for infections of humans and horses. Am J Trop Med Hyg. 1987;36(1):114-119. doi:10.4269/ajtmh.1987.36.114.

49. Mohamed M, McLees A, Elliott RM. Viruses in the Anopheles A, Anopheles B, and Tete Serogroups in the Orthobunyavirus Genus (Family Bunyaviridae) Do Not Encode an NSs Protein. J Virol. 2009;83(15):7612-7618. doi:10.1128/JVI.02080-08.

50. Williams JE, Imlarp S, Top FH, Cavanaugh DC, Russell PK. Kaeng Khoi virus from naturally infected bedbugs (Cimicidae) and immature free tailed bats. Bull World Health Organ. 1976;53(4):365-369.

51. Patroca da Silva S, Chiang JO, Marciel de Souza W, et al. Characterization of the Gamboa Virus Serogroup (Orthobunyavirus Genus, Peribunyaviridae Family). Am J Trop Med Hyg. 2018;98(5):tpmd170810. doi:10.4269/ajtmh.17-0810.

52. Plyusnin A, Elliott RM. Bunyaviridae: Molecular and Cellular Biology.; 2011.

53. World Health Organization. WHO Model List of Essential Medicines. 2017;Edition 14(March).

54. Ministry of Health. UGANDA CLINICAL GUIDELINES. 2012.

55. Li R, Faden DL, Fakhry C, et al. Clinical, genomic, and metagenomic characterization of oral tonguesquamous cell carcinoma in patients who do not smoke. Head Neck. 2014;36(10):1391. doi:10.1002/HED.

56. Stenglein MD, Sanchez-Migallon Guzman D, Garcia VE, et al. Differential Disease Susceptibilities in Experimentally Reptarenavirus-Infected Boa Constrictors and Ball Pythons. J Virol. 2017;91(15):e00451-17. doi:10.1128/JVI.00451-17.

57. Dobin A, Davis CA, Schlesinger F, et al. STAR: Ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15-21. doi:10.1093/bioinformatics/bts635.

58. Ruby JG, Bellare P, Derisi JL. PRICE: software for the targeted assembly of components of (Meta) genomic sequence data. G3 (Bethesda). 2013;3(5):865-880.

doi:10.1534/g3.113.005967.

59. Li W, Godzik A. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22(13):1658-1659. doi:10.1093/bioinformatics/btl158.

60. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357-359. doi:10.1038/nmeth.1923.

61. Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics. 2010;26(7):873-881. doi:10.1093/bioinformatics/btq057.

62. Ye Y, Choi J-H, Tang H. RAPSearch: a fast protein similarity search tool for short reads. BMC Bioinformatics. 2011;12(1):159. doi:10.1186/1471-2105-12-159.

63. Wilson MR, Fedewa G, Stenglein MD, et al. Multiplexed Metagenomic Deep Sequencing To Analyze the Composition of High-Priority Pathogen Reagents. mSystems. 2016;1(4):e00058-16. doi:10.1128/mSystems.00058-16.

64. Bogaert D, Keijser B, Huse S, et al. Variability and diversity of nasopharyngeal microbiota in children: A metagenomic analysis. PLoS One. 2011;6(2). doi:10.1371/journal.pone.0017035.

65. Pérez-Losada M, Alamri L, Crandall KA, Freishtat RJ. Nasopharyngeal microbiome diversity changes over time in children with asthma. PLoS One. 2017;12(1):1-13. doi:10.1371/journal.pone.0170543.

66. Hooper L V, Gordon JI. Commensal Host-Bacterial Relationships in the Gut. Science (80-). 2001;1115(10):1-7. doi:10.1126/science.1058709.

67. Sekirov I, Russell S, Antunes L. Gut microbiota in health and disease. Physiol Rev. 2010;90(3):859-904. doi:10.1152/physrev.00045.2009.

68. Bankevich A, Nurk S, Antipov D, et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J Comput Biol. 2012;19(5):455-477. doi:10.1089/cmb.2012.0021.

69. Mitchell A, Chang HY, Daugherty L, et al. The InterPro protein families database: The classification resource after 15 years. Nucleic Acids Res. 2015;43(D1):D213-D221. doi:10.1093/nar/gku1243.

70. Tsirigos KD, Peters C, Shu N, Käll L, Elofsson A. The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. Nucleic Acids Res. 2015;43(W1):W401-W407. doi:10.1093/nar/gkv485.

71. Giske CG, Cornaglia G. Supranational surveillance of antimicrobial resistance: The legacy of the last decade and proposals for the future. Drug Resist Updat. 2010;13(4-5):93-98. doi:10.1016/j.drup.2010.08.002.

Figure 1

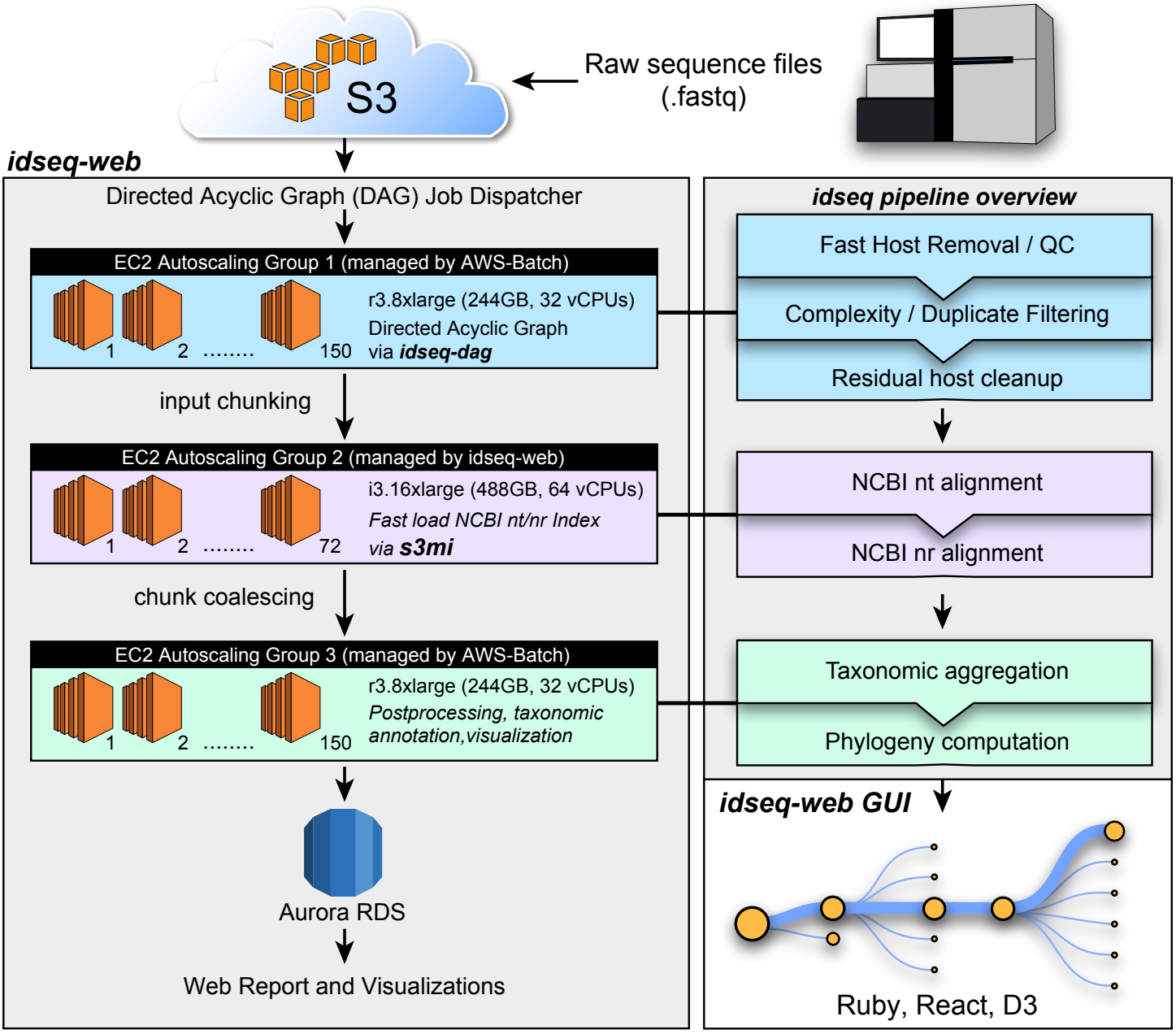
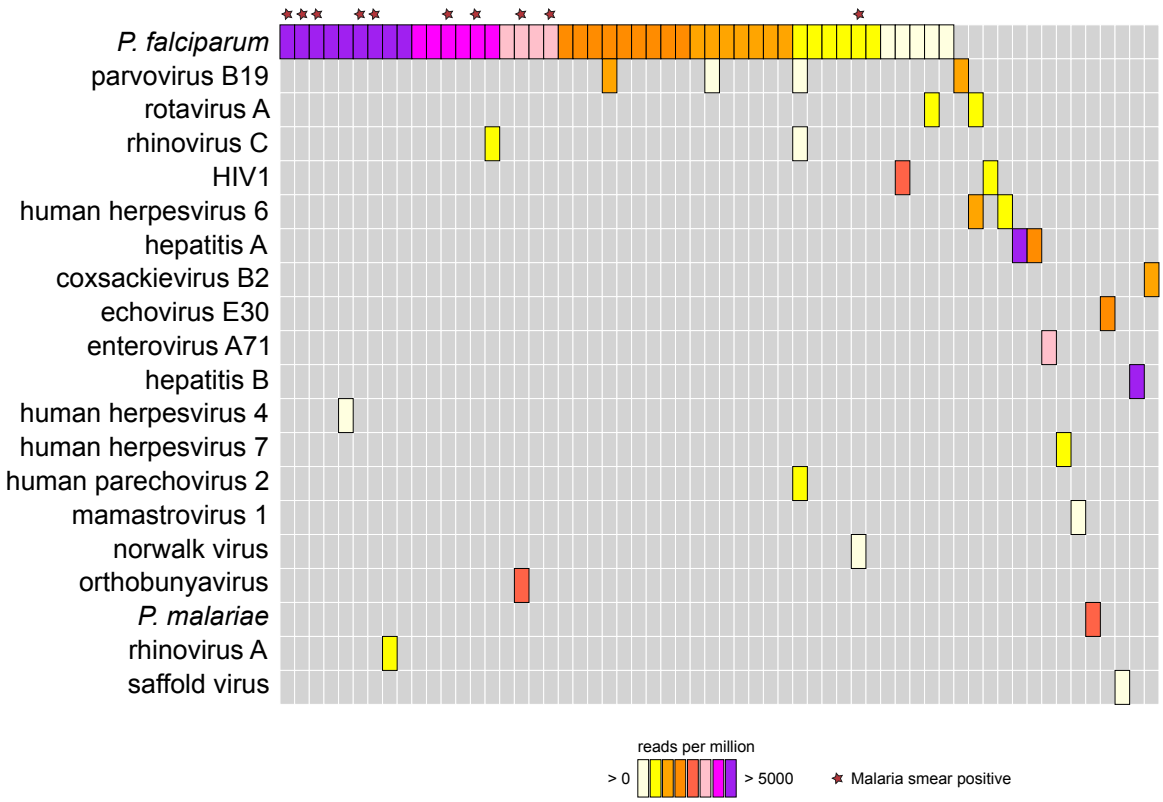

S3
Raw sequence files
(.fastq)
idseq-web
Directed Acyclic Graph (DAG) Job Dispatcher
EC2 Autoscaling Group 1 (managed by AWS-Batch)
1
2
.....
150
r3.8xlarge (244GB, 32 vCPUs)
Directed Acyclic Graph
via idseq-dag
input chunking
EC2 Autoscaling Group 2 (managed by idseq-web)
1
2
.....
72
i3.16xlarge (488GB, 64 vCPUs)
Fast load NCBI nt/nr Index
via s3mi
chunk coalescing
EC2 Autoscaling Group 3 (managed by AWS-Batch)
1
2
.....
150
r3.8xlarge (244GB, 32 vCPUs)
Postprocessing, taxonomic
annotation,visualization
Aurora RDS
Web Report and Visualizations
idseq pipeline overview
Fast Host Removal / QC
Complexity / Duplicate Filtering
Residual host cleanup
NCBI nt alignment
NCBI nr alignment
Taxonomic aggregation
Phylogeny computation
idseq-web GUI
Ruby, React, D3

Figure 2

A


P. falciparum
parvovirus B19
rotavirus A
rhinovirus C
HIV1
human herpesvirus 6
hepatitis A
coxsackievirus B2
echovirus E30
enterovirus A71
hepatitis B
human herpesvirus 4
human herpesvirus 7
human parechovirus 2
mamastrovirus 1
norwalk virus
orthobunyavirus
P. malariae
rhinovirus A
saffold virus
reads per million
> 0
> 5000
Malaria smear positive

B

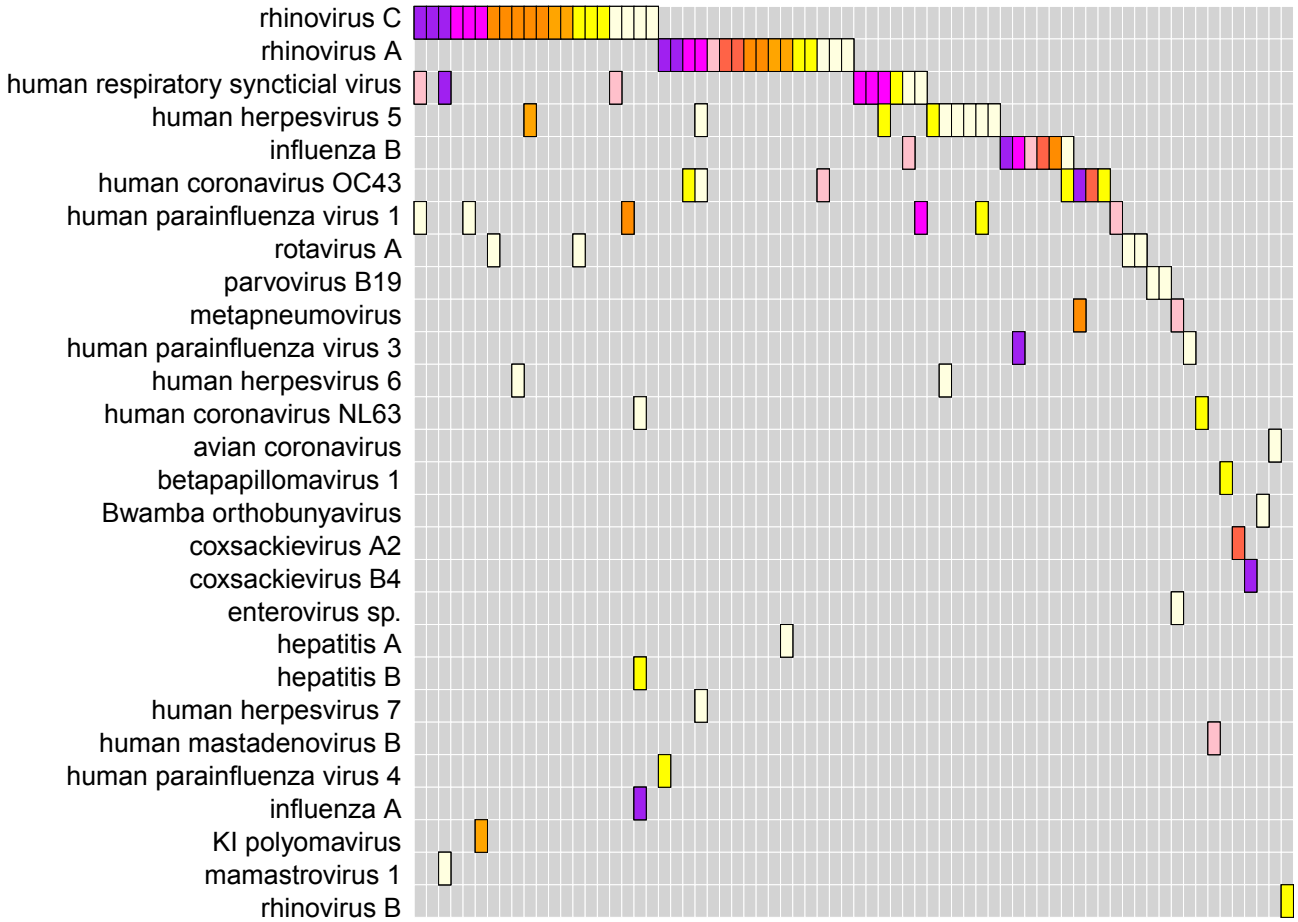

rhinovirus C
rhinovirus A
human respiratory syncticial virus
human herpesvirus 5
influenza B
human coronavirus OC43
human parainfluenza virus 1
rotavirus A
parvovirus B19
metapneumovirus
human parainfluenza virus 3
human herpesvirus 6
human coronavirus NL63
avian coronavirus
betapapillomavirus 1
Bwamba orthobunyavirus
coxsackievirus A2
coxsackievirus B4
enterovirus sp.
hepatitis A
hepatitis B
human herpesvirus 7
human mastadenovirus B
human parainfluenza virus 4
influenza A
KI polyomavirus
mamastrovirus 1
rhinovirus B

Figure 3

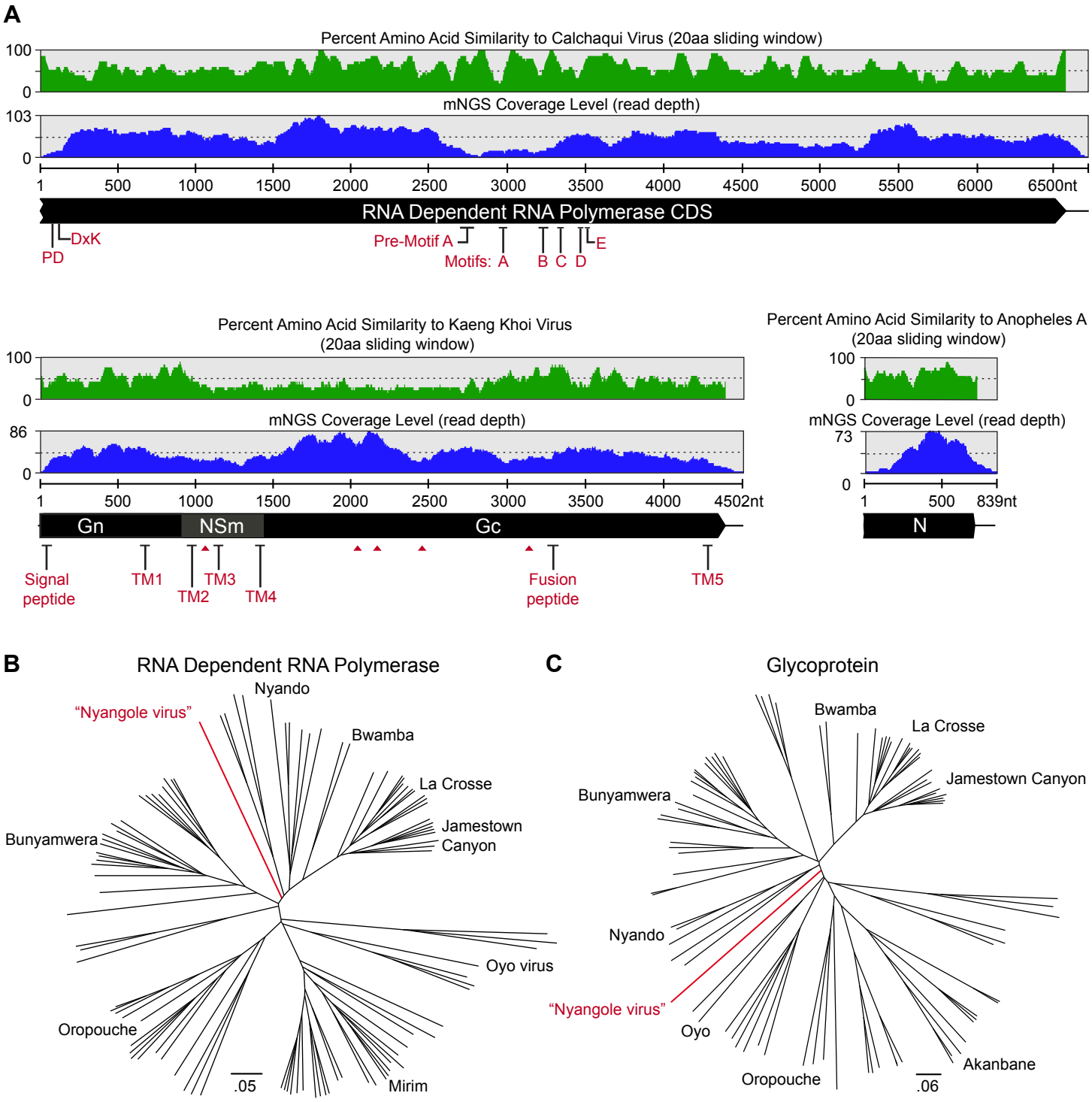

A
Percent Amino Acid Similarity to Calchaqui Virus (20aa sliding window)
100
0
mNGS Coverage Level (read depth)
103
0
1
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
5500
6000
6500nt
RNA Dependent RNA Polymerase CDS
PD
DxK
Pre-Motif A
Motifs: A
B
C
D
E
Percent Amino Acid Similarity to Kaeng Khoi Virus
(20aa sliding window)
100
0
mNGS Coverage Level (read depth)
86
0
1
500
1000
1500
2000
2500
3000
3500
4000
4502nt
Gn
NSm
Gc
Signal peptide
TM1
TM2
TM3
TM4
Fusion peptide
TM5
Percent Amino Acid Similarity to Anopheles A
(20aa sliding window)
100
0
mNGS Coverage Level (read depth)
73
0
1
500
839nt
N
B
RNA Dependent RNA Polymerase
Nyando
"Nyangole virus"
Bwamba
La Crosse
Jamestown
Canyon
Bunyamwera
Oyo virus
Oropouche
.05
Mirim
C
Glycoprotein
Bwamba
La Crosse
Jamestown Canyon
Bunyamwera
Nyando
"Nyangole virus"
Oyo
Oropouche
.06
Akanbane

Figure 4

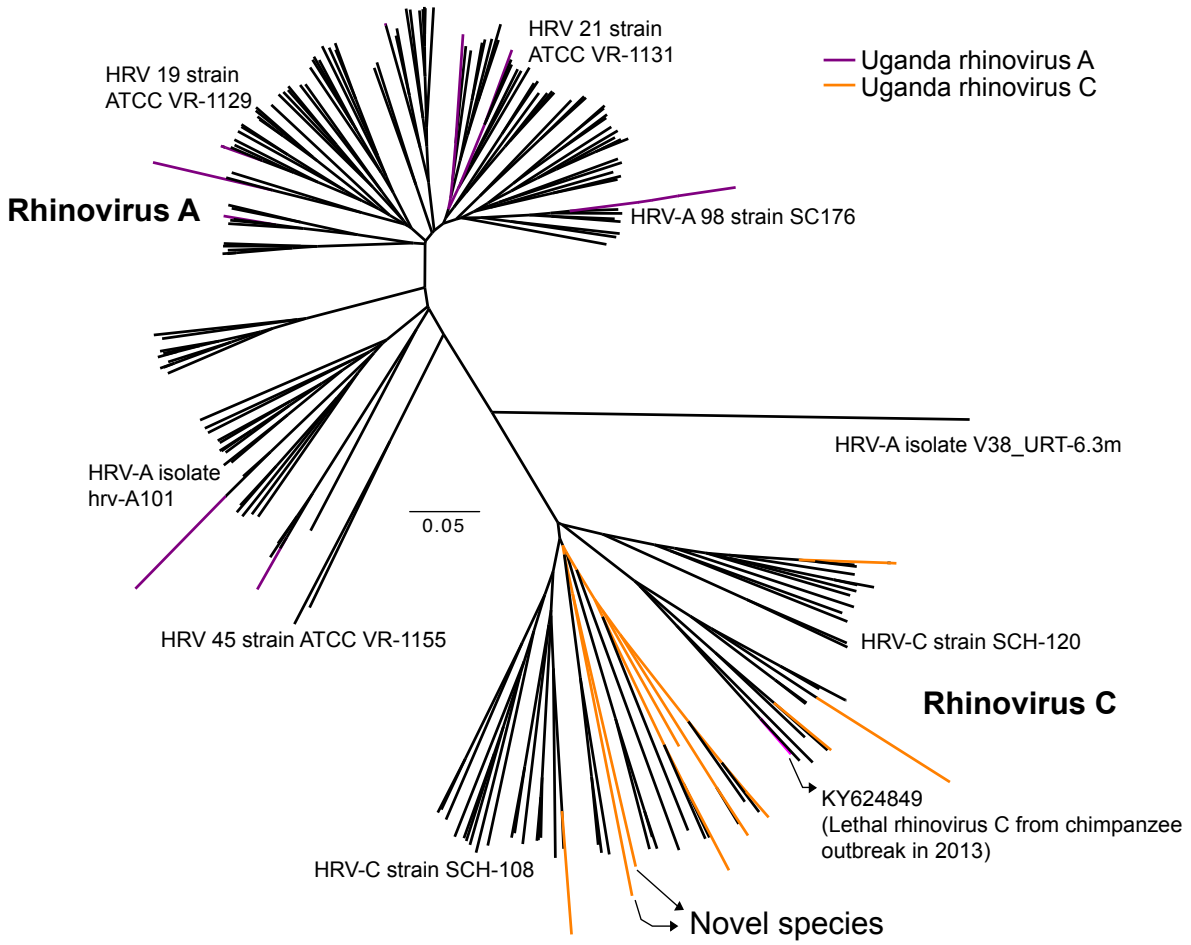

HRV 21 strain
ATCC VR-1131
Uganda rhinovirus A
Uganda rhinovirus C
HRV 19 strain
ATCC VR-1129
Rhinovirus A
HRV-A 98 strain SC176
HRV-A isolate V38_URT-6.3m
HRV-A isolate
hrv-A101
0.05
HRV 45 strain ATCC VR-1155
HRV-C strain SCH-120
Rhinovirus C
KY624849
(Lethal rhinovirus C from chimpanzee
outbreak in 2013)
HRV-C strain SCH-108
Novel species

Figure S1

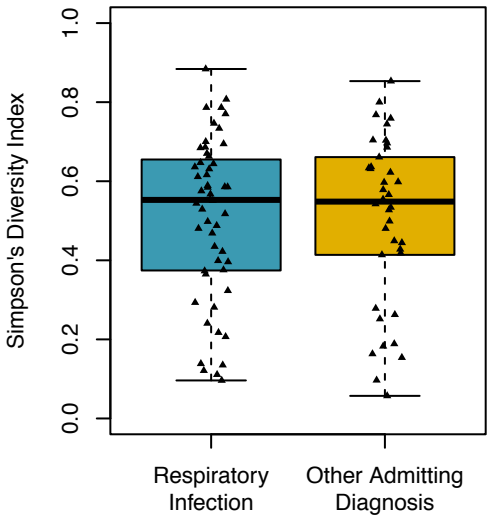
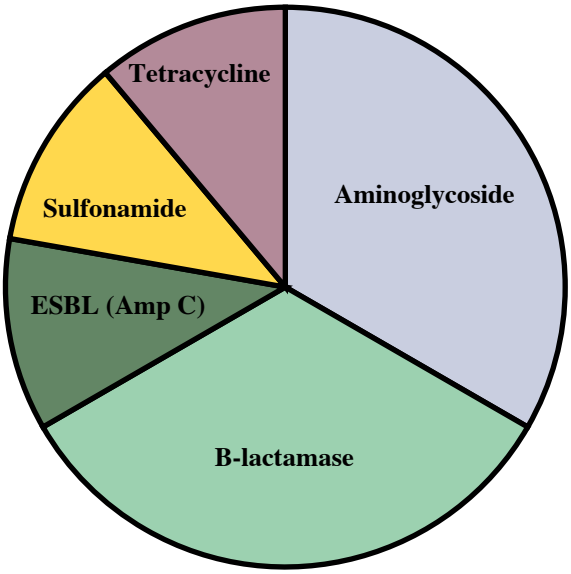

Simpson's Diversity Index
0.0
0.2
0.4
0.6
0.8
1.0
Respiratory Infection
Other Admitting Diagnosis

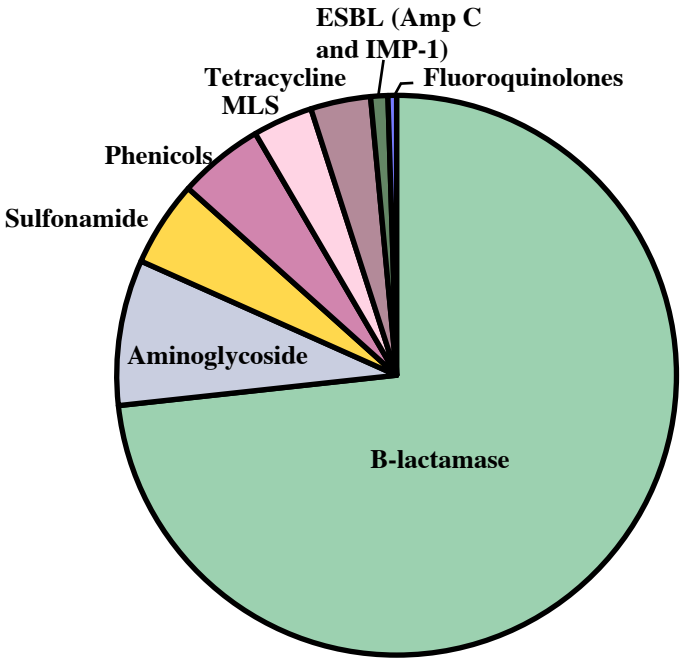
Figure S2

A



Category	AMR gene	Number
Aminoglycoside	StrA	3
B-lactamase	TEM-1D	3
ESBL	AmpC1	1
Sulfonamide	SulII	1
Tetracycline	TetO	1

B

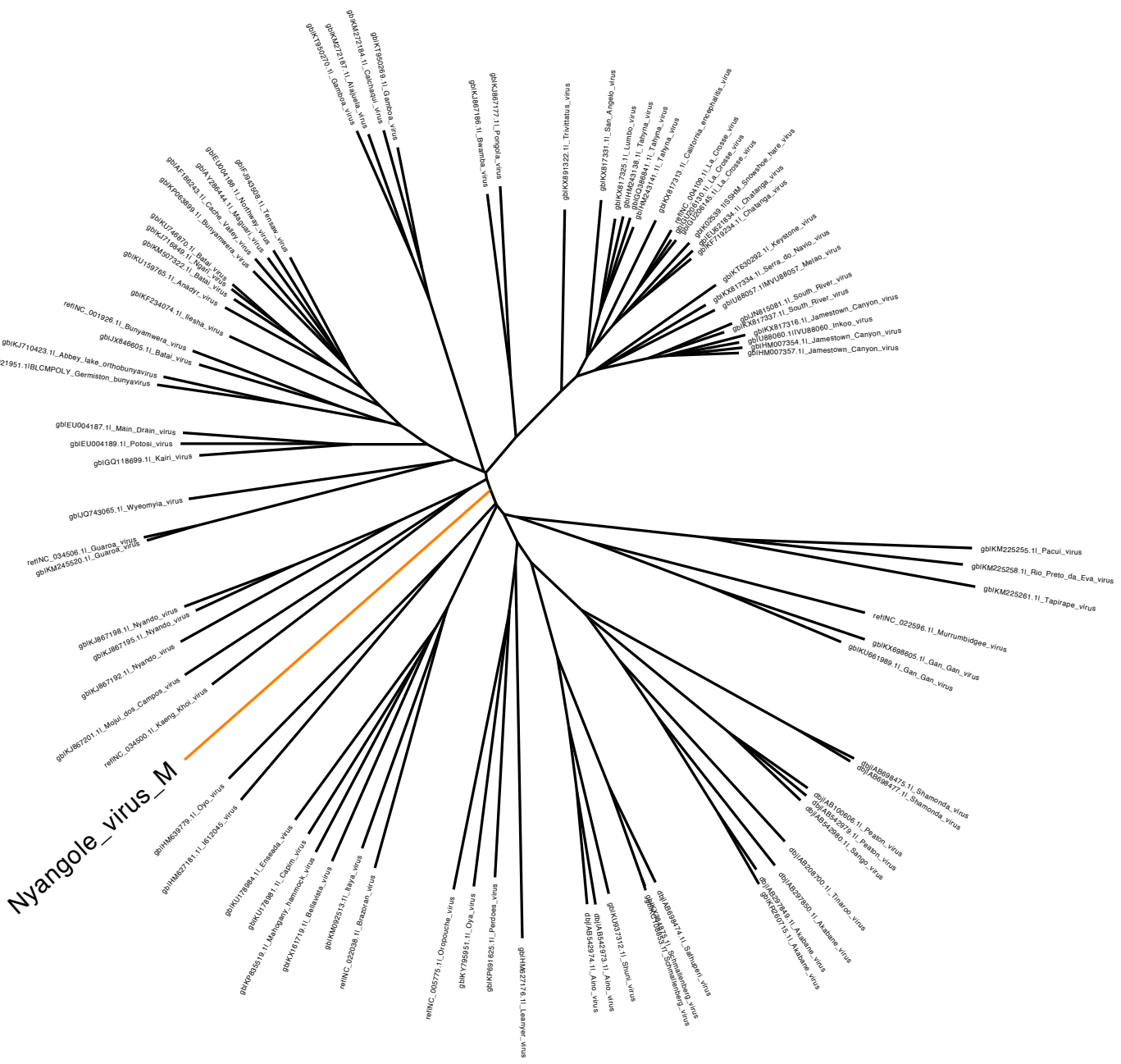


Category	AMR gene	Number
Aminoglycoside	StrA	8
Aminoglycoside	APH(3')-Ia	4
Aminoglycoside	StrB	3
Aminoglycoside	APH-Stph	1
Aminoglycoside	AacAad	1
B-lactamase	BRO	49
B-lactamase	PBP1b	46
B-lactamase	PBP1a	41
B-lactamase	TEM-1D	9
B-lactamase	CfxA	2
ESBL	AmpC1	1
ESBL	IMP-1	1
Fluoroquinolones	NorA	1
MLS	ErmX	4
MLS	MsrD	2
MLS	MphC	1
Phenicols	Cmr	4
Phenicols	CatA2	3
Phenicols	CatA9	1
Phenicols	CatBx	1
Sulfonamide	SulI	5
Sulfonamide	SulII	5
Tetracycline	TetB	3
Tetracycline	Tet-38	1
Tetracycline	TetC	1
Tetracycline	TetK	1
Tetracycline	TetM	1

Figure S3a

0.05

Figure S3b



0.06

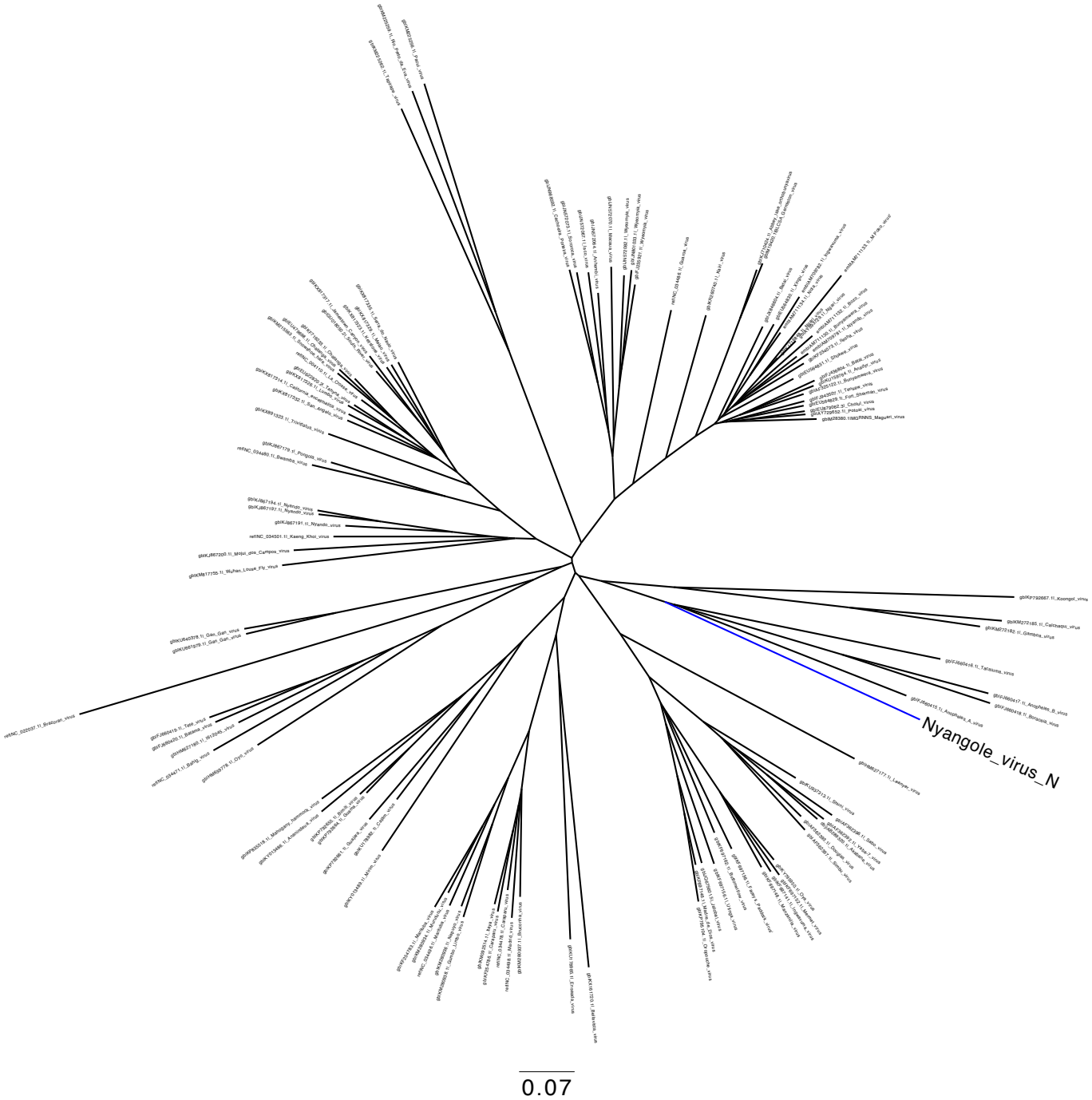



Figure S3c