Resource

A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples

Samia N. Naccache,[1,2] Scot Federman,[1,2] Narayanan Veeraraghavan,[1,2] Matei Zaharia,[3] Deanna Lee,[1,2] Erik Samayoa,[1,2] Jerome Bouquet,[1,2] Alexander L. Greninger,[4] Ka-Cheung Luk,[5] Barryett Enge,[6] Debra A. Wadford,[6] Sharon L. Messenger,[6] Gillian L. Genrich,[1] Kristen Pellegrino,[7] Gilda Grard,[8] Eric Leroy,[8] Bradley S. Schneider,[9] Joseph N. Fair,[9] Miguel A. Martínez,[10] Pavel Isa,[10] John A. Crump,[11,12,13] Joseph L. DeRisi,[4] Taylor Sittler,[1] John Hackett, Jr.,[5] Steve Miller,[1,2] and Charles Y. Chiu[1,2,14,15]

[1]Department of Laboratory Medicine, UCSF, San Francisco, California 94107, USA; [2]UCSF-Abbott Viral Diagnostics and Discovery Center, San Francisco, California 94107, USA; [3]Department of Computer Science, University of California, Berkeley, California 94720, USA; [4]Department of Biochemistry, UCSF, San Francisco, California 94107, USA; [5]Abbott Diagnostics, Abbott Park, Illinois 60064, USA; [6]Viral and Rickettsial Disease Laboratory, California Department of Public Health, Richmond, California 94804, USA; [7]Department of Family and Community Medicine, UCSF, San Francisco, California 94143, USA; [8]Viral Emergent Diseases Unit, Centre International de Recherches Médicales de Franceville, Franceville, BP 769, Gabon; [9]Metabiota, Inc., San Francisco, California 94104, USA; [10]Departamento de Genética del Desarrollo y Fisiología Molecular, Instituto de Biotecnología, Universidad Nacional Autónoma de México, Cuernavaca, Morelos, 62260, Mexico; [11]Division of Infectious Diseases and International Health and the Duke Global Health Institute, Duke University Medical Center, Durham, North Carolina 27708, USA; [12]Kilimanjaro Christian Medical Centre, Moshi, Kilimanjaro, 7393, Tanzania; [13]Centre for International Health, University of Otago, Dunedin, 9054, New Zealand; [14]Department of Medicine, Division of Infectious Diseases, UCSF, San Francisco, California 94143, USA

Unbiased next-generation sequencing (NGS) approaches enable comprehensive pathogen detection in the clinical microbiology laboratory and have numerous applications for public health surveillance, outbreak investigation, and the diagnosis of infectious diseases. However, practical deployment of the technology is hindered by the bioinformatics challenge of analyzing results accurately and in a clinically relevant timeframe. Here we describe SURPI ("sequence-based ultrarapid pathogen identification"), a computational pipeline for pathogen identification from complex metagenomic NGS data generated from clinical samples, and demonstrate use of the pipeline in the analysis of 237 clinical samples comprising more than 1.1 billion sequences. Deployable on both cloud-based and standalone servers, SURPI leverages two state-of-the-art aligners for accelerated analyses, SNAP and RAPSearch, which are as accurate as existing bioinformatics tools but orders of magnitude faster in performance. In *fast* mode, SURPI detects viruses and bacteria by scanning data sets of 7–500 million reads in 11 min to 5 h, while in *comprehensive* mode, all known microorganisms are identified, followed by de novo assembly and protein homology searches for divergent viruses in 50 min to 16 h. SURPI has also directly contributed to real-time microbial diagnosis in acutely ill patients, underscoring its potential key role in the development of unbiased NGS-based clinical assays in infectious diseases that demand rapid turnaround times.

[Supplemental material is available for this article.]

There is great interest in the use of unbiased next-generation sequencing (NGS) technology for comprehensive detection of pathogens from clinical samples (Dunne et al. 2012; Wylie et al. 2012; Chiu 2013; Firth and Lipkin 2013). Conventional diagnostic testing for pathogens is narrow in scope and fails to detect the etiologic agent in a significant percentage of cases (Barnes et al. 1998; Louie et al. 2005; van Gageldonk-Lafeber et al. 2005; Bloch and Glaser 2007; Denno et al. 2012). Failure to accurately diagnose and treat infection in a timely fashion contributes to continued transmission and increased mortality in hospitalized patients (Kollef et al. 2008). Ongoing discovery of novel pathogens, such as Bas-Congo rhabdovirus (Grard et al. 2012) and MERS (Middle East Respiratory Syndrome) coronavirus (Zaki et al. 2012), also underscores the need for rapid, broad-spectrum diagnostic assays that are able to recognize these emerging agents.

[15]Corresponding author
E-mail charles.chiu@ucsf.edu
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.171934.113. Freely available online through the *Genome Research* Open Access option.

24:1180–1192 Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/14; www.genome.org

Unbiased NGS holds the promise of identifying all potential pathogens in a single assay without a priori knowledge of the target. Given sufficiently long read lengths, multiple hits to the microbial genome, and a well-annotated reference database, nearly all microorganisms can be uniquely identified on the basis of their specific nucleic acid sequence. Thus, NGS has widespread microbiological applications, including infectious disease diagnosis in clinical laboratories (Dunne et al. 2012), pathogen discovery in acute and chronic illnesses of unknown origin (Chiu 2013), and outbreak investigation on a global level (Firth and Lipkin 2013). However, the latest NGS laboratory workflows incur minimum turnaround times exceeding 8 h from clinical sample to sequence (Quail et al. 2012). Thus, it is critical that subsequent computational analyses of NGS data be performed within a timeframe suitable for actionable responses in clinical medicine and public health (i.e., minutes to hours). Such pipelines must also retain sensitivity, accuracy, and throughput in detecting a broad range of clinically relevant pathogenic microorganisms.

Computational analysis of metagenomic NGS data for pathogen identification remains challenging for several reasons. First, alignment/classification algorithms must contend with massive amounts of sequence data. Recent advances in NGS technologies have resulted in instruments that are capable of producing >100 gigabases (Gb) of reads in a day (Loman et al. 2012). Reference databases of host and pathogen sequences range in size from 2 Gb for viruses to 3.1 Gb for the human genome to 42 Gb for all nucleotide sequences in the National Center for Biotechnology Information (NCBI) nucleotide (nt) collection (NCBI nt DB) as of January 2013. Second, only a small fraction of short NGS reads in clinical metagenomic data typically correspond to pathogens (a "needle-in-a-haystack" problem) (Kostic et al. 2012; Wylie et al. 2012; Yu et al. 2012), and such sparse reads often do not overlap sufficiently to permit de novo assembly into longer contiguous sequences (contigs) (Kostic et al. 2011). Thus, individual reads, typically only 100–300 nucleotides (nt) in length, must be classified to a high degree of accuracy. Finally, novel microorganisms with divergent genomes, particularly viruses, are not adequately represented in existing reference databases and often can only be identified on the basis of remote amino acid homology (Xu et al. 2011; Grard et al. 2012).

To address these challenges, the most widely used approach is computational subtraction of reads corresponding to the host (e.g., human), followed by alignment to reference databases that contain sequences from candidate pathogens (MacConaill and Meyerson 2008; Greninger et al. 2010; Kostic et al. 2011; Zhao et al. 2013). Traditionally, the BLAST algorithm (Altschul et al. 1990) is used for classification of human and nonhuman reads at the nucleotide level (BLASTn), followed by low-stringency protein alignments using a translated nucleotide query (BLASTx) for detection of divergent sequences from novel pathogens (Delwart 2007; Briese et al. 2009; Xu et al. 2011; Grard et al. 2012; Chiu 2013). However, BLAST is too slow for routine analysis of NGS metagenomics data (Niu et al. 2011), and end-to-end processing times, even on multicore computational servers, can take several days to weeks. Analysis pipelines that use faster, albeit less sensitive, algorithms upfront for host computational subtraction, such as PathSeq (Kostic et al. 2011), still rely on traditional BLAST approaches for final pathogen determination. In addition, whereas PathSeq works well for tissue samples in which the vast majority of reads are host-derived and thus subject to subtraction, the pipeline becomes computationally prohibitive when analyzing complex clinical metagenomic samples open to the environment, such as respiratory secretions or stool (Fig. 1B; Supplemental Table S1). Other published pipelines are focused solely on limited detection of specific types of microorganisms, are unable to identify highly divergent novel pathogens, and/or utilize computationally taxing algorithms such as BLAST (Bhaduri et al. 2012; Borozan et al. 2012; Dimon et al. 2013; Naeem et al. 2013; Wang et al. 2013; Zhao et al. 2013). Furthermore, there is hitherto scarce reported data on the real-life performance of these pipelines for pathogen identification in clinical samples.

Here we describe SURPI ("sequence-based ultrarapid pathogen identification"), a cloud-compatible bioinformatics analysis pipeline that provides extensive classification of reads against viral and bacterial databases in *fast* mode and against the entire NCBI nt DB in *comprehensive* mode (Fig. 1A). Novel pathogens are also identified in *comprehensive* mode by amino acid alignment to viral and/or NCBI nr protein databases. Notably, SURPI generates results in a clinically actionable timeframe of minutes to hours by leveraging two alignment tools, SNAP (Fig. 1C; Zaharia et al. 2011) and RAPSearch (Fig. 1D; Zhao et al. 2012), which have computational times that are orders of magnitude faster than other available algorithms. Here we evaluate the performance of these tools for pathogen detection using both in silico-generated and clinical data and describe use of the SURPI pipeline in the analysis of 15 independent NGS data sets consisting of 157 clinical samples multiplexed across 47 barcodes and including over 1.1 billion reads. These data sets encompass a variety of clinical infections, detected pathogens, sample types, and depths of coverage. We also demonstrate use of the pipeline for detection of emerging novel outbreak viruses and for clinical diagnosis of a case of unknown fever in a returning traveler.

Results

Accuracy of SURPI aligners (SNAP and RAPSearch) using in silico data

The accuracy of SURPI was evaluated by benchmarking its nucleotide alignment tool, SNAP, against BLASTn and two other aligners commonly used for human genome mapping, BWA (Li and Durbin 2009) and Bowtie 2 (BT2) (Fig. 2A–E; Langmead and Salzberg 2012). In addition, SURPI's protein similarity search tool, RAPSearch (Zhao et al. 2012), was directly compared to BLASTx (Fig. 2F; Altschul et al. 1990). A query data set of 100 base pair (bp) reads was randomly generated in silico from human, bacterial, and viral reference databases. The data set consisted of 1 million human reads, 250,000 bacterial reads, 25,000 viral reads, and 1000 reads each from four known viruses (norovirus, ebolavirus, human immunodeficiency virus [HIV-1], and influenza A), and three divergent "novel" viruses whose genomes had been removed a priori from the reference database (Supplemental Table S2): Bas-Congo rhabdovirus (BASV) (Grard et al. 2012), titi monkey adenovirus (TMAdV) (Chen et al. 2011), and bat influenza H17N10 (Tong et al. 2012). Receiver operating characteristic (ROC) curves (Akobeng 2007) were generated to assess the sensitivity and specificity of each aligner in classifying human, bacterial, or viral reads. All nucleotide aligners shared >99.5% optimal sensitivity and specificity for human sequence identification (Fig. 2A), with SNAP exhibiting the highest specificity (>99.8%) and comparable sensitivity to BLASTn (99.9% versus 100%). For bacterial detection (Fig. 2B), SNAP was more accurate than BWA and BT2, and exhibited reduced sensitivity (99.5%) albeit superior specificity (98.5%) relative to BLASTn (100% and 97.9%), as was also the trend for viral detection (Fig. 2C). The accuracy of all four tools in identifying sequences from

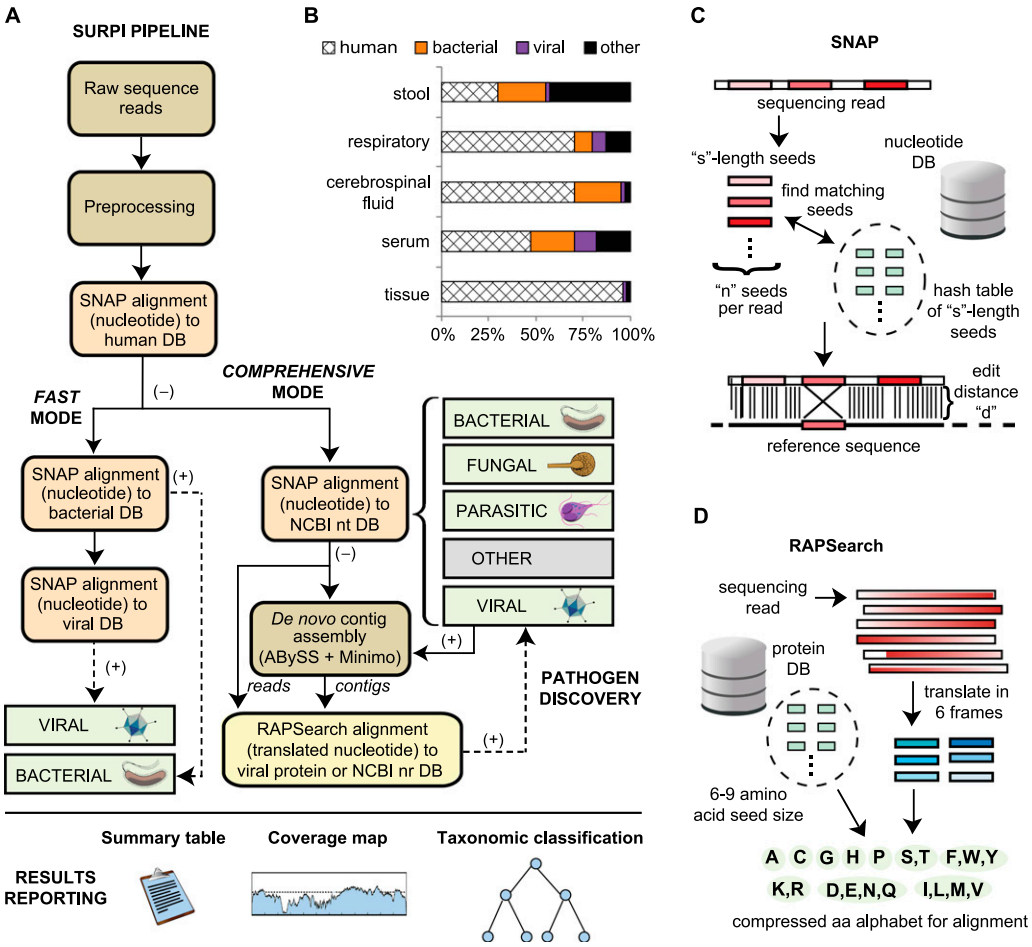

Figure 1. The SURPI pipeline for pathogen detection. (*A*) A schematic overview of the SURPI pipeline. Raw NGS reads are preprocessed by removal of adapter, low-quality, and low-complexity sequences, followed by computational subtraction of human reads using SNAP. In *fast* mode, viruses and bacteria are identified by SNAP alignment to viral and bacterial nucleotide databases. In *comprehensive* mode, reads are aligned using SNAP to all nucleotide sequences in the NCBI nt collection, enabling identification of bacteria, fungi, parasites, and viruses. For pathogen discovery of divergent microorganisms, unmatched reads and contigs generated from de novo assembly are then aligned to a viral protein database or all protein sequences in the NCBI nr collection using RAPSearch. SURPI reports include a list of all classified reads with taxonomic assignments, a summary table of read counts, and both viral and bacterial genomic coverage maps. (*B*) Relative proportion of NGS reads classified as human, bacterial, viral, or other in different clinical sample types. (*C*) The SNAP nucleotide aligner (Zaharia et al. 2011). SNAP aligns reads by generating a hash table of sequences of length "s" from the reference database and then comparing the hash index with "n" seeds of length "s" generated from the query sequence, producing a match based on the edit distance "d." (*D*) The RAPSearch protein similarity search tool (Zhao et al. 2012). RAPSearch aligns translated nucleotide queries to a protein database using a compressed amino acid alphabet at the level of chemical similarity for greatly increased processing speed.

known viruses was comparable (Fig. 2D), but SNAP and BLASTn were superior to BWA and BT2, in identifying reads from divergent viruses using low-stringency parameters (Fig. 2E). Nevertheless, the overall poor performance of all four nucleotide aligners in detecting divergent viral reads (<20% sensitivity) underscored the need for translated nucleotide alignment algorithms such as RAPSearch and BLASTx (Briese et al. 2009; Xu et al. 2011; Grard et al. 2012; Swei et al. 2013). By ROC curve analysis, these two algorithms performed similarly in the detection of sequences from divergent viral genomes (Fig. 2F).

Speed of SURPI aligners (SNAP and RAPSearch) using in silico data

The computational speed of SNAP relative to BLASTn, BT2, and BWA in aligning NGS reads to the human hg19 database (human DB) was evaluated using progressively larger in silico query data sets of 1.25 million, 25 million, 125 million, and 1.25 billion reads (Fig. 3A; Supplemental Tables S2, S3). BLASTn alignments were associated with prohibitively long run times, consuming >19 h to analyze only 1.25 million reads, with proportionally longer estimated times for the larger data sets. Although all three remaining aligners performed comparably well with the 1.25 million read data set, SNAP scaled significantly better with larger data sets and was 23–87× faster than BWA and BT2.

Next, we investigated the feasibility of using the SNAP algorithm to align reads to all sequences in the 42 Gb NCBI nt DB. Computational subtraction of human host sequences followed by SNAP alignment to the entire NCBI nt DB was accomplished in under 1 h for 1.25 million and <40 h for 1.25 billion reads (Fig. 3B,C). Overall timing metrics for SURPI, whether using a cloud server or local server, were comparable (Fig. 3B), likely due to the use of high-performance, low-latency solid-state drives (SSDs) for the cloud (Supplemental Methods). We also benchmarked the

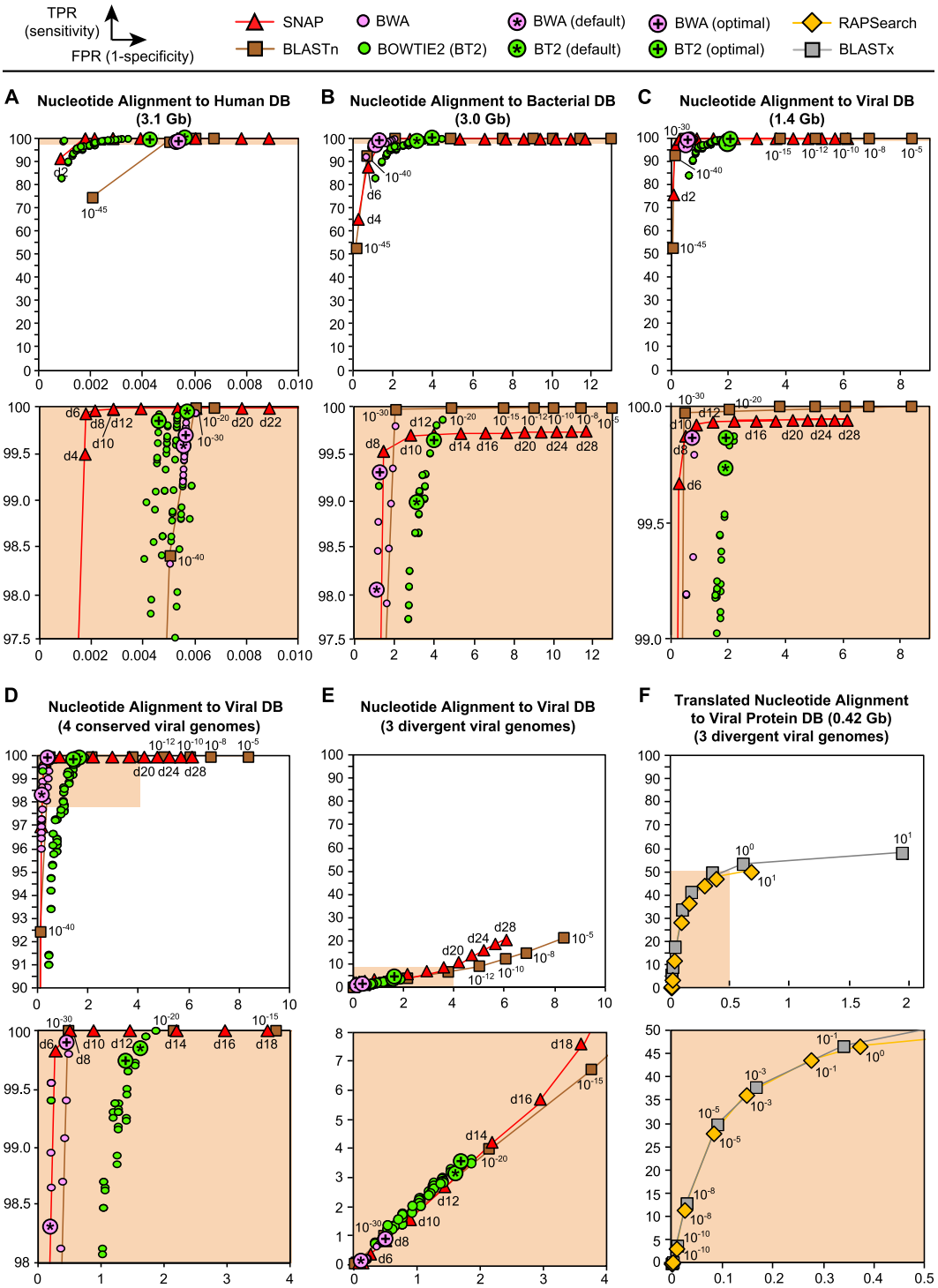

Figure 2. SURPI aligners (SNAP and RAPSearch) are comparable to other tested aligners for detection of human, bacterial, and viral reads from in silico-generated query data sets. ROC curves were generated to evaluate the ability of four nucleotide aligners (SNAP, BWA, BT2, and BLASTn) to correctly detect in silico-generated NGS reads when mapped against the human DB (*A*), bacterial DB (*B*), or viral nucleotide DB (*C*). The accuracy of detection was assessed using Youden's index and the F_1 score. Sensitivity or the true positive rate (TPR) (*y*-axis) is plotted against 1-specificity or the false positive rate (FPR) (*x*-axis). (*D*) Detection of reads corresponding to four viral genomes [norovirus, Zaire ebolavirus, influenza A(H1N1)pdm09, and HIV-1] by nucleotide alignment. (*E*) Detection of reads corresponding to three divergent viruses (TMAdV, BASV, and bat influenza H17N10, a novel influenza strain) by nucleotide alignment. (*F*) Detection of reads corresponding to three divergent viruses (TMAdV, BASV, and bat influenza H17N10) by translated nucleotide (protein) alignment using the RAPSearch and BLASTx aligners. The sequences of these viruses were removed from the nucleotide and protein viral reference databases prior to alignment. The *lower* shaded panels are magnifications of the corresponding shaded boxed regions in the *upper* panels.

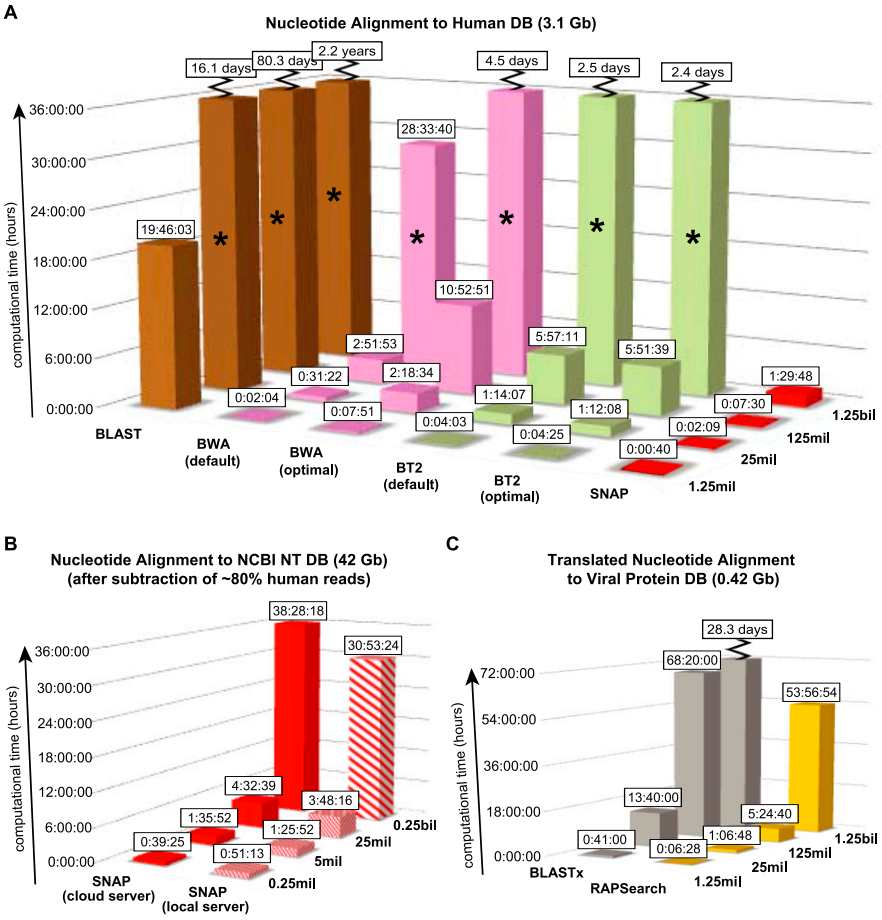

Figure 3. SURPI aligners (SNAP and RAPSearch) are significantly faster than other tested aligners and scale better with larger data sets. Timing performance was benchmarked on a single computational server using in silico query data sets of increasing size. The breaks (zigzag lines) represent computational times that are off-scale. Some of the computational times were estimated (asterisks). (*A*) Performance time for alignment of reads to the human DB. (*B*) Performance time for SNAP alignment of reads to the entire 42-Gb NCBI nt DB. The *z*-axis denotes the approximate number of remaining reads following computational subtraction against the human DB. SNAP performance times were benchmarked separately on local and cloud servers. (*C*) Performance times for translated nucleotide alignment to the viral protein DB using RAPSearch and BLASTx.

speed of RAPSearch relative to BLASTx in aligning translated query reads to a viral protein database (viral protein DB). RAPSearch was found to be 5–10× faster than BLASTx across all query data sets (Fig. 3C).

Accuracy of the SURPI aligners (SNAP and RAPsearch) using clinical sample data

To evaluate the "real-life" performance of SNAP relative to BLASTn, BT2, and BWA, and of RAPSearch relative to BLASTx, ROC curves for viral detection were generated from three computationally challenging NGS data sets (Fig. 4A–C; Supplemental Table S3). The query sets corresponded to (1) a complex metagenomic stool sample pool (Yu et al. 2012) harboring a respiratory syncytial virus (RSV) strain with ~10% genomic sequence divergence, (2) a nasal swab sample pool from patients infected with 2009 pandemic influenza H1N1 [influenza A(H1N1)pdm09] and sequenced using short 65-bp reads (Greninger et al. 2010), and (3) a serum sample from a patient with hantavirus pulmonary syndrome from Sin Nombre virus (SNV) infection (Nunez et al. 2014) and sequenced using longer 250-bp reads. Both SNAP and BLASTn exhibited superior sensitivity than BWA and BT2 in detection of reads corresponding to these three viruses. Across all three data sets, 100% specificity was retained using an expectation value (E-value) cutoff of 1×10^{-15} for BLASTn and an edit distance of 12 for SNAP. At that threshold cutoff, the sensitivities of SNAP and BLASTn for detection of each virus were similar (84.3%/85.1% for RSV, 99.6%/98.7% for influenza A(H1N1)pdm09, and 93.8%/99.7% for hantavirus). Among the 15 true hantavirus reads not detected by SNAP and accounting for the reduced 93.8% sensitivity, 10 were found to be chimeric reads, while five were reads with internal regions of low-quality data. Cropping the long 250-bp reads in the hantavirus data set to 75 bp improved sensitivity from 93.8% to 98.1% due to increased detection of these previously undetected reads without affecting specificity (Fig. 4C). By ROC curve analysis, RAPSearch had comparable accuracy to BLASTx across all three clinical data sets (Fig. 4A–C, bottom panels).

The combined in silico and clinical data on SNAP and RAPSearch performance (Figs. 2–4) guided (1) the choice of an edit distance "d" of 12 as the most appropriate empirical cutoff for SNAP alignment (Fig. 1C), (2) read cropping prior to SNAP alignment to a length of 75 bp to maximize sensitivity for chimeric reads or reads with error-prone 3′ ends from Illumina sequencing and to allow use of a fixed edit distance threshold (Fig. 4C; Supplemental Results; Supplemental Fig. S1), and (3) the serial coupling of the SNAP and RAPSearch algorithms to maximize speed without sacrificing breadth of detection.

Accurate detection of pathogens from clinical samples using SURPI

SURPI was used to accurately classify viral pathogens down to the species and even strain level in various clinical metagenomic NGS data sets (Fig. 5A–G; Supplemental Results; Supplemental Tables S4, S5), automatically generating summary tables (Supplemental Tables S6–S21) and coverage maps (Supplemental Fig. S2) corresponding to the actual virus present in the sample. Plasma samples spiked with human immunodeficiency virus (HIV-1) at titers ranging from 10^2 to 10^4 copies/mL were identified and mapped to the correct strain (Fig. 5A), showing a linear correlation between number of aligned reads and viral titer, while sapovirus (SaV) and a divergent human parechovirus 1 (HPeV1) shed in children with diarrhea and provisionally named HPeV-1 isolate MX1 were correctly identified (Fig. 5B), as was human herpesvirus 3 (HHV3) in cerebrospinal fluid (CSF) from a patient

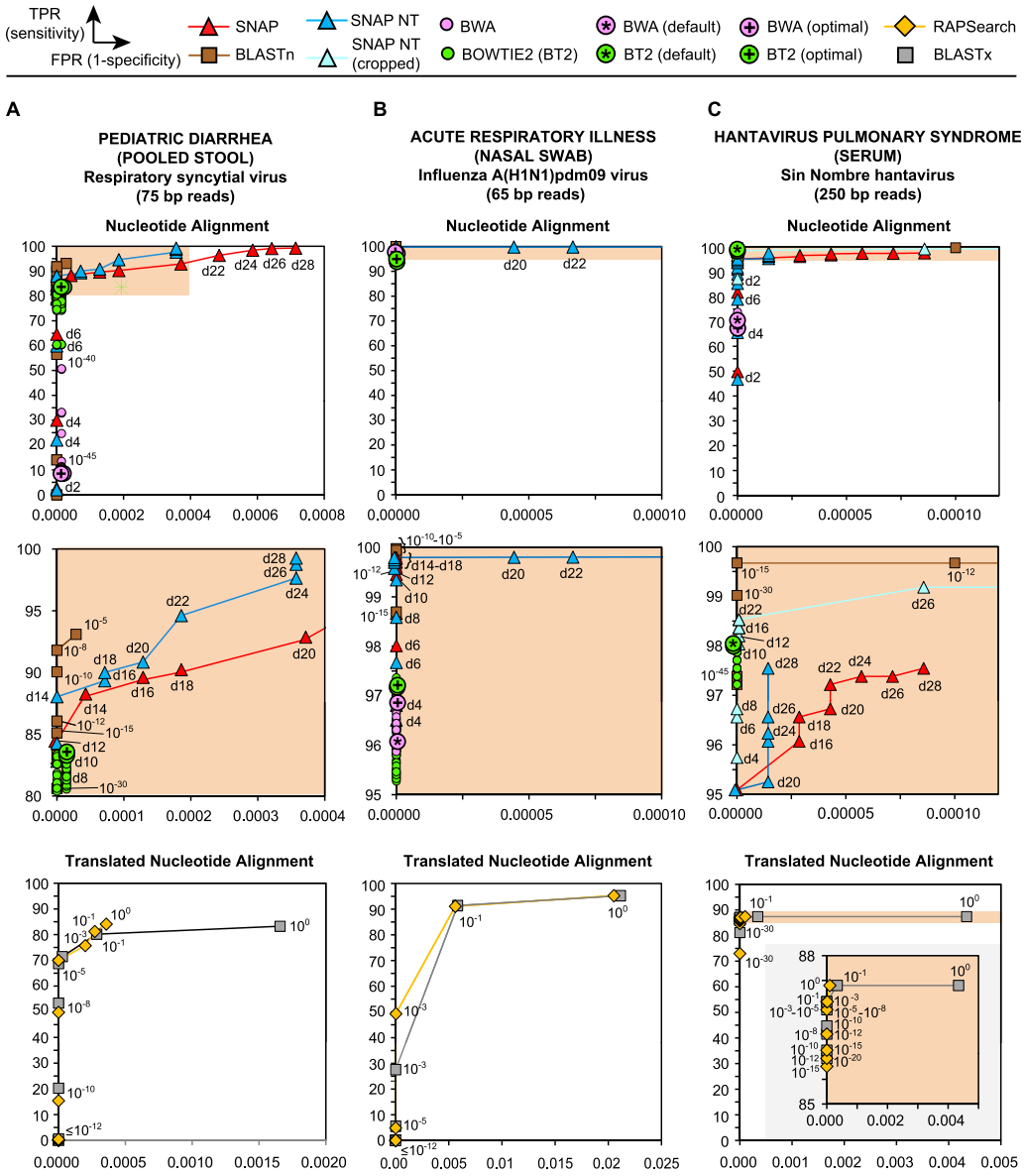

Figure 4. SURPI aligners (SNAP and RAPSearch) are comparable to other tested aligners for detection of viral reads in clinical NGS data sets. ROC curves were generated to evaluate the ability of nucleotide and translated nucleotide (protein) aligners to detect reads corresponding to three target viruses: (*A*) respiratory syncytial virus (RSV) from stool; (*B*) influenza A(H1N1)pdm09 from a nasal swab; and (*C*) Sin Nombre hantavirus from serum. Sensitivity or the true positive rate (TPR) (*y*-axis) is plotted against 1-specificity or the false positive rate (FPR) (*x*-axis). For each aligner, reads assigned to the correct viral genus were used for generating the ROC curve. The shaded panels are magnifications of the corresponding shaded regions in the *upper* panels (*A–C*, nucleotide alignment) or overlapping larger panel (*C*, translated nucleotide alignment).

with encephalitis (Fig. 5D), and hepatitis C virus subtype 1b (HCV-1b) from a patient with transfusion-transmitted hepatitis (Fig. 5E). SURPI also successfully identified human papillomavirus 18 (HPV-18) in sequences from an infected prostate cancer cell line used in the literature as a benchmarking standard (Fig. 5C), and produced nearly identical contigs to those generated by three other computational pipelines (Kostic et al. 2011; Bhaduri et al. 2012; Naeem et al. 2013). In *comprehensive* mode, incorporating both nucleotide alignments to the NCBI nt DB and viral protein similarity searches, SURPI was able to detect essentially 100% of viral reads corresponding to a given species (Fig. 5A–E). In contrast, due to lack of a protein similarity search step, SURPI in *fast* mode showed noticeably less coverage for more divergent viruses such as the HPeV-1 isolate MX1 in diarrheal stool (Fig. 5B), which shared <80% nucleotide identity with its closest related genome in the reference database (Supplemental Table S4). In addition to sequences from known pathogens, bacteriophage sequences derived from commensal bacteria and/or reagent and laboratory contamination were common and detected in nearly all clinical NGS data sets.

In comprehensive mode, SURPI was also able to identify highly divergent novel viruses whose sequences had been removed a priori from the reference database (Fig. 5F,G; Supplemental Table S4). The genome of TMAdV, an adenovirus causing a fulminant pneumonia outbreak in a titi monkey colony in California with cross-species infection of a human researcher

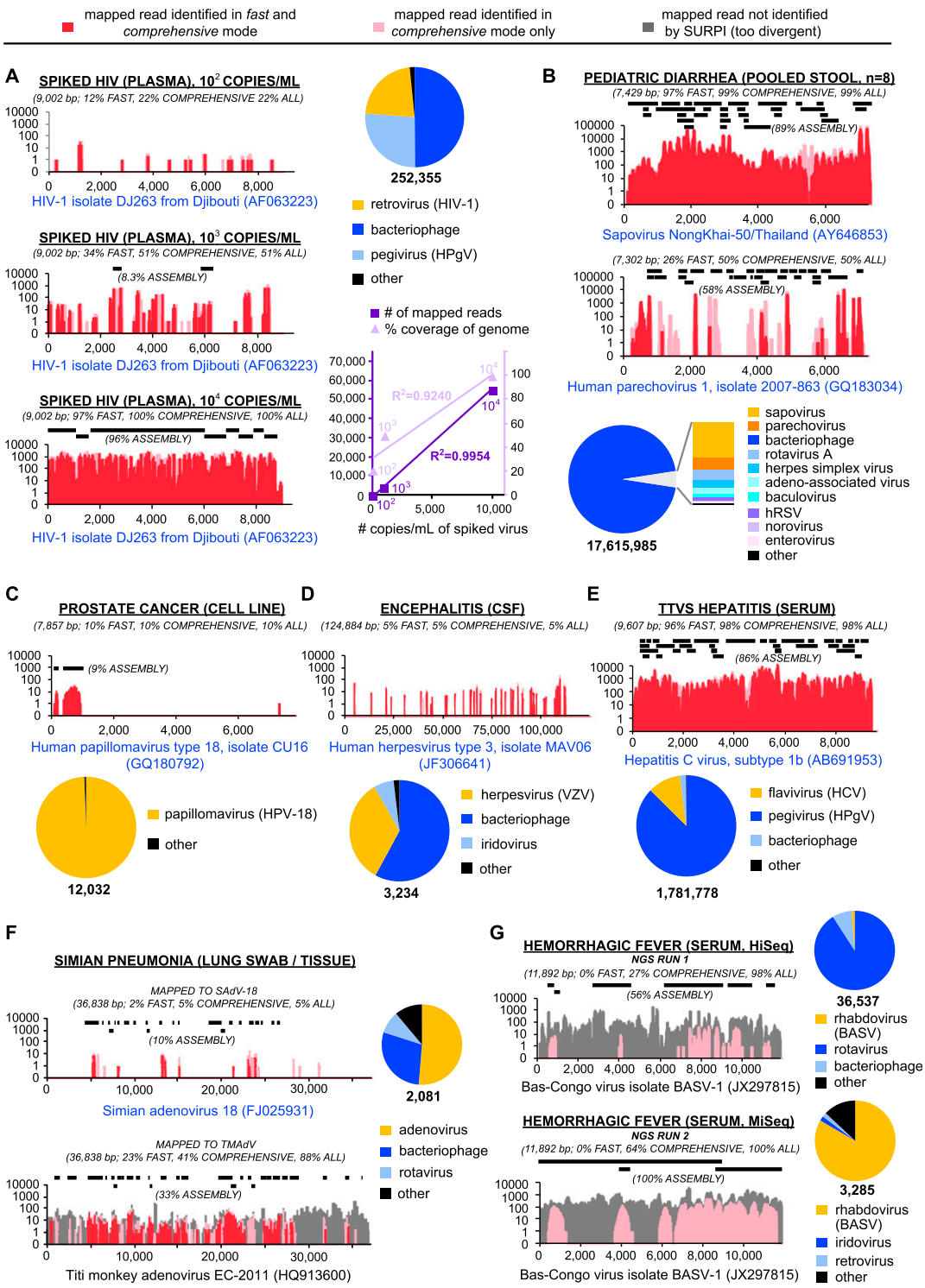

Figure 5. The SURPI pipeline correctly identifies viral species in clinical NGS data sets. Data sets corresponding to clinical samples or sample pools harboring target viral pathogens were analyzed using SURPI. Pie charts show detected viruses derived from the output summary tables. Target viruses are color-coded in yellow or orange; other viruses are color-coded ranked by their relative abundance in shades of blue, followed by shades of purple. Coverage maps of the "best hit" viral genome in *fast* mode (red) and *comprehensive* mode (pink, overlaid by red) display automated SURPI output corresponding to the detected target viral genome (blue text). The read coverage (*y*-axis, log scale) and de novo assembled contigs (black lines) are plotted as a function of nucleotide position along the genome (*x*-axis). Percent coverage achieved using SURPI in *fast* mode ("*FAST*"), in *comprehensive* mode ("*COMPREHENSIVE*"), and by de novo assembly ("*ASSEMBLY*"), as well as the actual coverage from all reads in the data set ("*ALL*") are shown. (*A*) Coverage plots of HIV-1 spiked at titers of 10^2–10^4 copies/mL. The number of mapped reads and percent coverage are plotted against the viral copy number (*inset*). Coverage plots of SaV and HPeV-1 (*B*), HPV-18 (*C*), HHV-3 (*D*), and HCV-1b (*E*). (*F*) Coverage plot mapping SURPI-classified genus-level *Mastadenovirus* reads (red/pink) to the SAdV-18 genome, or *Mastadenovirus* reads (red/pink) and all specific TMAdV reads (gray) to the TMAdV genome. (*G*) Coverage plots mapping SURPI-classified family-level *Rhabdoviridae* reads (pink) or all specific BASV reads (gray) to the BASV genome.

(Chen et al. 2011), shares <50% amino acid identity with its closest known relative, simian adenovirus 18 (SAdV-18). Nevertheless, SURPI, in either *fast* or *comprehensive* mode, successfully identified adenovirus sequences in lung samples from moribund titi monkeys with TMAdV pneumonia, and mapped these reads to the genome of SAdV-18 (Fig. 5F). In *comprehensive* mode, coverage of 41% of the TMAdV genome was achieved. For the hemorrhagic fever-associated BASV rhabdovirus (Grard et al. 2012), sharing <34% amino acid identity to its closest relative, SURPI failed to detect any rhabdovirus reads at the nucleotide level in *fast* mode, whereas in *comprehensive* mode, reads classified as *Rhabdoviridae* were detected on the basis of protein homology using RAPSearch (Fig. 5G). As laboratory contamination by rotavirus was noted in the previously published NGS data set (Grard et al. 2012), a subsequent NGS data set was generated from the same serum aliquot. This data set was devoid of rotaviral sequences, and because of longer reads and higher coverage of BASV (Supplemental Tables S4–S5), enabled de novo assembly of 100% of the viral genome (Fig. 5G, lower panel). Notably, a two-tiered de novo assembly approach involving the use of both de Bruijn

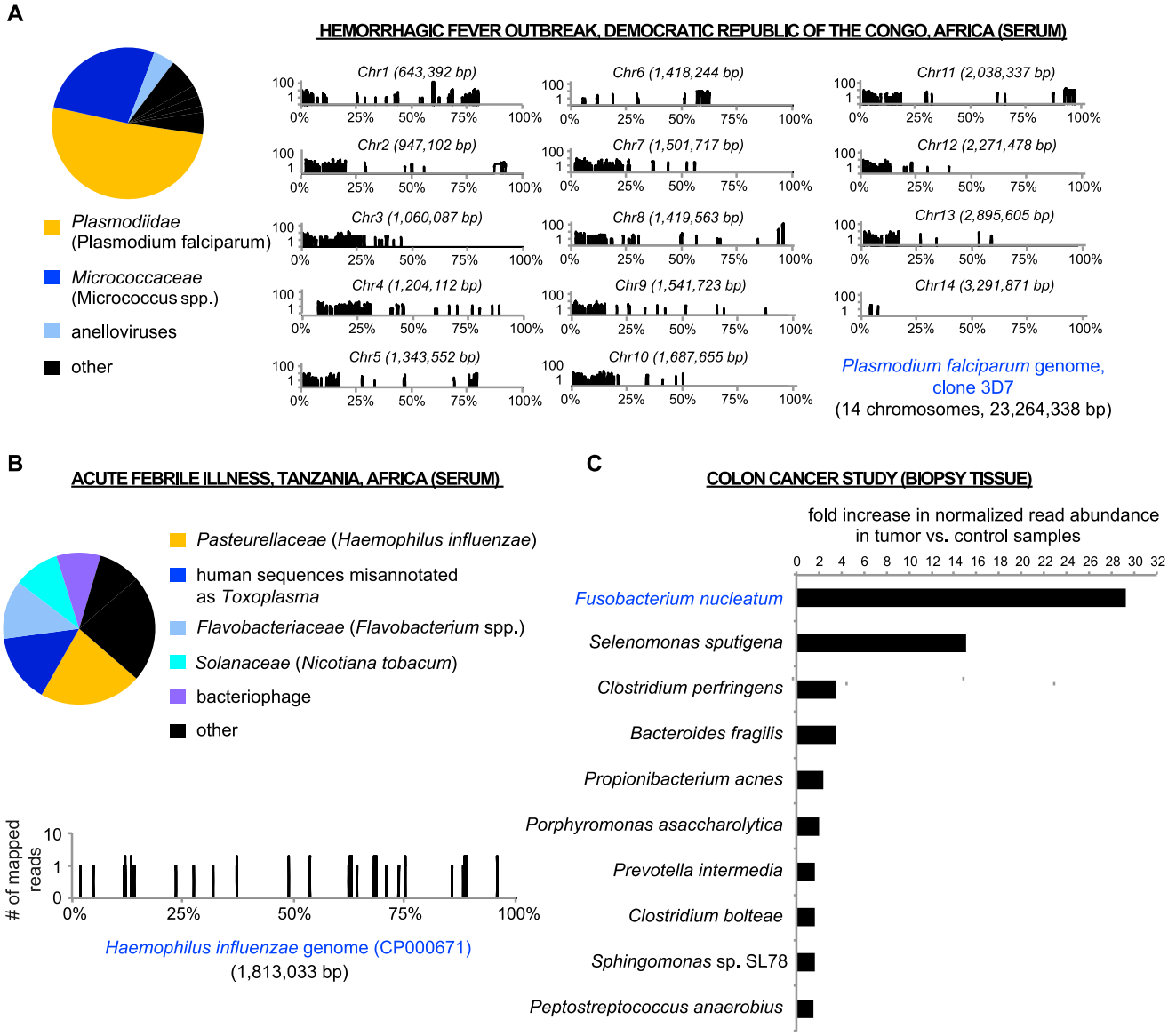

Figure 6. The SURPI pipeline correctly identifies bacterial and parasitic species in clinical NGS data sets. Three NGS data sets corresponding to clinical samples or sample pools and found to harbor target pathogenic bacteria or parasites were analyzed using SURPI in *comprehensive* mode. Pie charts represent the breakdown of SURPI-classified pathogen reads by family. (*A*) Serum from an individual with acute hemorrhagic fever in the Democratic Republic of the Congo (DRC), Africa, was analyzed by unbiased NGS. NGS reads identified as *Plasmodium* by SURPI are mapped to the 14 chromosomes of *Plasmodium falciparum* clone 3D7, including multiple hits to telomeric ends by reads corresponding to the *var* gene (Gardner et al. 2002). (*B*) Serum from a patient who died from a critical febrile illness in Tanzania, Africa (Crump et al. 2013) was analyzed using NGS. SURPI generates a coverage map corresponding to the "best hit" bacterial genome, *Haemophilus influenza*e. (*C*) SURPI was used to classify the diversity of bacterial species in 22 clinical samples, 11 from colorectal tumors and 11 from normal tissue (Castellarin et al. 2012). For the top 10 bacterial species, the fold-increase in the average normalized abundance between normal and diseased tissue is plotted in rank order from most to least abundant.

(Simpson et al. 2009) and overlap-layout-consensus (OLC) (Sommer et al. 2007) methods produced longer and higher quality assemblies than the use of de Bruijn algorithms alone (Supplemental Methods; Supplemental Table S22).

Detection of pathogens by SURPI was not restricted to viruses. SURPI was able to identify NGS reads spanning all 14 chromosomes from the malarial parasite *Plasmodium falciparum* in a hemorrhagic fever patient from Gabon, Africa (Fig. 6A; Gardner et al. 2002). The bacterial pathogen *Haemophilus influenzae* was found in serum from a patient enrolled in a febrile illness cohort from Tanzania, Africa (Fig. 6B; Crump et al. 2013). In addition, SURPI was used to analyze publicly available data from a study that detected an increase in bacterial reads aligning to *Fusobacterium nucleatum* in association with colorectal tumors (Castellarin et al. 2012). As in the previous study, an overabundance of aligned reads corresponding to *Fusobacteirum nucleatum* was identified, with a 29-fold increase in tumor versus control samples (Fig. 6C).

Speed of SURPI and feasibility for real-time clinical analysis

To benchmark end-to-end processing times for SURPI across a range of clinical diseases and samples, we analyzed NGS metagenomic data corresponding to 15 data sets ranging from 6.7 to 509 million reads in size (Fig. 7A; Supplemental Fig. S3). In *fast* mode, processing times ranged from 11 min to nearly 5 h and were linearly proportional to the number of reads (R^2 = 0.9742) (Fig. 7A, inset) , while in *comprehensive* mode, SURPI took 59 min to 16 h (R^2 = 0.4473) (Fig. 7A, inset). Overall, SURPI in *fast* mode was ~5.3× faster than in *comprehensive* mode, running at 18,700 versus 3500 reads/sec, respectively, but a greater proportion of reads were classified by SURPI in *comprehensive* mode (Supplemental Fig. S3).

Unbiased NGS and SURPI analyses were applied in a clinical setting to analyze an acute serum sample from a 20-yr-old female patient presenting with 3 d of fever to 101.5°C, myalgias, and headache (Fig. 7B). The patient had just returned from hiking in

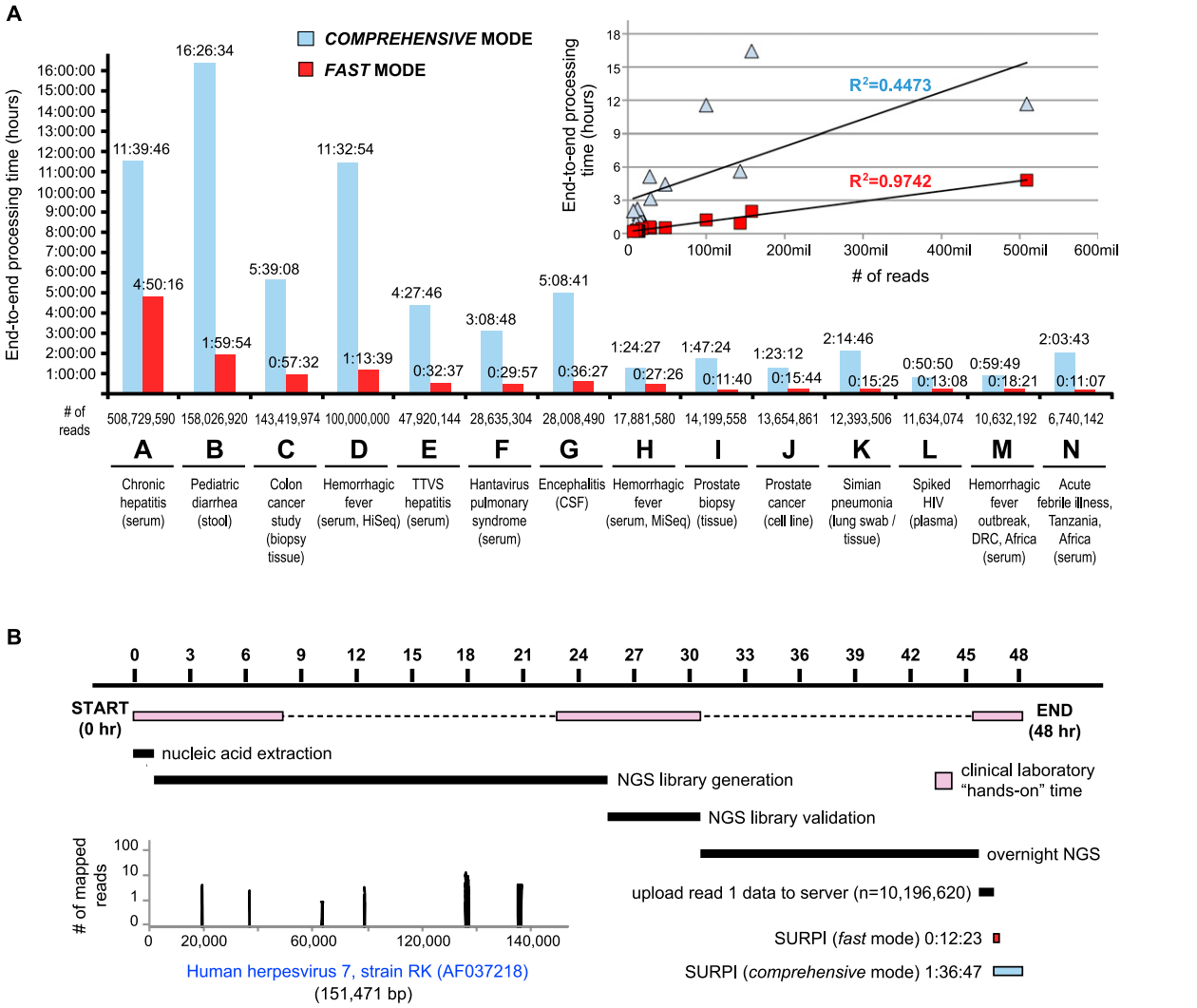

Figure 7. Speed of SURPI and feasibility for real-time clinical analysis. (*A*) Timing performance for SURPI in *fast* mode (red) and *comprehensive* mode (blue) was benchmarked on a single computational server across 12 NGS data sets representing a variety of infectious diseases and sample types. Processing end-to-end-times are plotted against the number of reads (*inset*), along with regression trend lines corresponding to SURPI processing in *fast* and *comprehensive* modes. (*B*) A serum sample from a returning traveler with an acute febrile illness was analyzed using NGS, resulting in SURPI detection of human herpesvirus 7 (HHV-7) infection (*inset*, coverage plot) in a clinically relevant 48-h timeframe.

a region of Australia endemic for the mosquito-borne Ross River and Barmah Forest alphaviruses (Knope et al. 2013). Within a 48-h sample-to-answer turnaround time and 13 min SURPI analysis time, sequences spanning the genome of human herpesvirus 7 (HHV-7) were detected. As no other pathogens were identified in patient sera and acute antibody IgG/IgM titers for HHV-7 were negative, the NGS results and subsequent confirmatory PCR supported a diagnosis of primary HHV-7 infection (Ward et al. 2002). The patient recovered spontaneously without any complications.

Discussion

Rapid improvements in NGS technology and widespread availability have sparked increased demand in clinical and public health settings. The MiSeqDX instrument has recently become the first next-generation sequencer approved by the FDA (Collins and Hamburg 2013), opening the door for routine implementation of NGS-based assays in the clinical laboratory. The speed of SURPI makes the pipeline well-suited for real-time clinical applications such as infectious disease diagnosis and outbreak response. In *fast* mode, end-to-end processing times for 7–50 million reads are ~10–30 min and reliably proportional to the size of the NGS data set (Fig. 7A), while in *comprehensive* mode, all potential pathogens (viruses, bacteria, fungi, and parasites) as well as novel emerging viruses with high sequence divergence can be identified in ~1–5 h. SURPI also includes an option to use the entire NCBI nonredundant protein collection (NCBI nr DB) for the protein alignment step (Fig. 1; Supplemental Fig. S4). In addition, SURPI can efficiently handle NGS data generated from complex metagenomic samples such as stool and respiratory secretions, which are exposed to the environment and contain a large proportion of non-host sequences. For example, PathSeq (Kostic et al. 2011) takes >7 d to identify all pathogens in a 150-million-read stool data set, whereas SURPI in *comprehensive* mode only uses 16 h of processing time (Supplemental Results; Supplemental Table S1). The pipeline also has the capacity to incorporate new reference databases as needed and permits potential quantitative or semiquantitative assessments of pathogen titer by retention of duplicate reads during SNAP analysis (Fig. 5A). Finally, SURPI combines de Bruijin and OLC de novo assembly methods in a two-tiered approach to generate longer contigs for identification of divergent viral sequences (Supplemental Table S22).

To our knowledge, SURPI is the only NGS pipeline for pathogen identification to be extensively tested across multiple clinical sample types representing a variety of infectious diseases (Figs. 5–7). SURPI runs as a single Linux script and configuration file and is available via direct download or as a self-deployable cloud computing instance. In particular, the speed and throughput of the pipeline make it highly convenient and cost-effective when run on a cloud server (Supplemental Table S1). In addition to maintaining the software, we are currently implementing further improvements to SURPI including development of a user-friendly graphical interface. SURPI analysis is also being incorporated as part of an ongoing effort to validate unbiased NGS assays for pathogen diagnosis in a CLIA (Clinical Laboratory Improvement Amendments)-certified laboratory. This will require constructing accurate, well-annotated reference databases and validating thresholds for the number and distribution of reads mapping to the genome to determine whether low-level detection of a known pathogen is clinically significant.

There is heightened awareness of the threat from both emerging and re-emerging infectious diseases and the need for enhanced surveillance to avert pandemics (Morse et al. 2012). In this study, SURPI accurately detected viral pathogens of realized or potential outbreak importance in clinical NGS data sets, including influenza A(H1N1)pdm09, TMAdV, BASV, and Sin Nombre hantavirus. At the same time, there is an urgent need for the implementation of rapid, highly multiplexed assays for infectious disease diagnosis in the microbiology laboratory. SURPI was used here to prospectively identify, in <13 min of NGS analysis time, HHV-7 infection in a returning traveler with fever. Although the negative acute IgG/IgM serologies are suggestive, we were unable to definitively establish whether this case represented primary infection or reactivation of HHV-7, given that convalescent sera was not available. However, NGS data from the same run was also analyzed using SURPI for actionable diagnosis of neuroleptospirosis in an immunocompromised child with a life-threatening meningoencephalitis, which dramatically impacted his treatment and resulted in a clinically favorable outcome (Wilson et al. 2014). Although many technical and regulatory challenges remain (Dunne et al. 2012; Gargis et al. 2012), bioinformatics analysis is no longer the weak link when deploying NGS as a clinical diagnostic tool for infectious diseases.

Methods

Clinical NGS data sets

Details regarding clinical samples, approved research protocols for sample collection, and NGS library construction are provided in the Supplemental Methods.

Hardware

Minimum hardware requirements for running SURPI include a multicore server running Ubuntu 12.04 (preferred) with at least 60 GB of RAM. SURPI and its software dependencies require ~1 GB of disk space. Reference data requires ~1 TB of disk space. During SURPI runtime, up to 10× the size of the input FASTQ file may be needed as additional temporary storage. The specific hardware used here for SURPI testing and benchmarking are provided in the Supplemental Methods.

Custom modifications to the SNAP nucleotide aligner

SNAP is a new, hash-based nucleotide aligner developed for mapping of NGS data to reference genomes across a wide range of read lengths (50–10,000 bp) (Zaharia et al. 2011). Available at http://snap.cs.berkeley.edu, SNAP runs 10–100× faster than existing tools while maintaining comparable or higher accuracy (Fig. 1C). The aligner partially derives its speed from loading the entire indexed reference database into RAM. Since SNAP was originally designed only for human genome (hg19) mapping, we generated a custom build tailored for alignment to different reference databases containing thousands of similar and/or overlapping sequences, such as bacterial RefSeq (Pruitt et al. 2007). This modified SNAP build (v0.15) included options to improve alignment speed and efficiency by stopping at the first hit ("-f" parameter) and retaining hits that mapped to multiple locations ("-x" parameter). All of the SNAP alignments used by SURPI incorporate these two additional parameters.

Reference databases

A description of how the reference databases used by SURPI were generated is given in the Supplemental Methods.

ROC curve analysis of in silico-generated query data sets

ROC curve analysis (Zweig and Campbell 1993) was used to evaluate the ability of the various nucleotide aligners (SNAP, BWA, BT2, and BLASTn) to correctly classify a given set of in silico NGS reads when mapped against the human DB, bacterial DB, or viral nucleotide DB (Fig. 2A–E). Similarly, ROC curve analysis was used to compare RAPSearch and BLASTx performance when mapping translated nucleotide reads to the viral protein DB (Fig. 2F). Details on the construction of the in silico query nucleotide data sets and range of parameters used for the ROC curve analysis (Supplemental Fig. S4) are provided in the Supplemental Methods. The gold standard criterion used for a correctly classified read to any given database was that the in silico read had originated from that database. To generate the ROC curves, the true positive rate [TPR = TP/(TP + FN)], or sensitivity, was plotted against the false positive rate [FPR = FP/(FP + TN)], or 1-specificity. Youden's index and the F_1 score (harmonic mean) were applied as independent criteria to select an optimal cutoff point of diagnostic accuracy for the ROC curve (Akobeng 2007). In all instances, the cutoff point identified by Youden's index and the F_1 score were identical.

Speed benchmarking for aligners

Details of speed benchmarking for the various alignment algorithms are provided in the Supplemental Methods.

ROC curve analysis of clinical query data sets

The ROC curve analysis used NGS data sets corresponding to a stool sample from a child in Mexico with diarrhea (Yu et al. 2012), a nasal swab sample from a patient with acute respiratory illness (Greninger et al. 2010), and a serum sample from a patient in California with hantavirus pulmonary syndrome (Nunez et al. 2014), harboring RSV, influenza A(H1N1)pdm2009, and Sin Nombre virus, respectively (Supplemental Methods). Seven million unique preprocessed reads were selected from each data set, and human and bacterial reads were removed prior to ROC curve analysis by SNAP alignment to the human DB and bacterial DB, respectively, using an edit distance of 12. The gold standard criterion for a correct viral classification was BLASTn alignment against the target viral genome (obtained by Sanger sequencing) at an E-value cutoff of 10^{-8}.

SURPI pipeline

The SURPI pipeline is comprised of a series of shell, Python, and Perl scripts in Linux and incorporates several open-source tools, including the SNAP and RAPSearch aligners. SURPI has a set of fixed external software and database dependencies (Supplemental Fig. S5) and user-defined custom parameters (Supplemental Methods). The pipeline accepts a raw FASTQ file as input and recognizes the presence of multiple barcodes used for indexing. Paired-end reads are handled by concatenating the files corresponding to the individual reads and their mate pairs into a single file for streamlined analysis. The preprocessing step consists of (1) trimming low-quality and adapter sequences using cutadapt (Martin 2011), retaining reads of trimmed length >50 bp, (2) removing low-complexity sequences using the DUST algorithm in PRINSEQ (Schmieder and Edwards 2011), and (3) normalizing read lengths for SNAP alignment by cropping reads of length >75 to 75 bp. In *fast* mode, SNAP alignments are first performed against the human DB followed by separate alignments of the human background-subtracted reads to bacterial and viral nucleotide DBs, whereas in *comprehensive* mode, the initial SNAP alignment against the human database is followed by sequential alignments to 29 indexed nt subdatabases. Following SNAP alignment, matched reads are taxonomically classified by lookup of matched GI/accession numbers from the NCBI taxonomy database. The taxonomic classification is then appended to the SAM (sequence alignment/map) file outputted by SNAP.

In *comprehensive* mode, the SURPI pipeline continues to the de novo assembly step, which uses an empiric approach that is optimized for NGS metagenomics data (Supplemental Material). Duplicates at the level of cropped reads are first removed using GenomeTools (gt) SEQUNIQ (Gremme et al. 2013). The corresponding full-length reads are then de-multiplexed by barcode and analyzed using the Message Passing Interface (MPI)-based parallel version of the AbySS de novo assembler (AbYSS 1.3.5 release) (Simpson et al. 2009). Increased robustness of the de Brujin graph-based assembly is obtained by running AbySS multiple times at a kmer size of 34, using both the entire data set and individually partitioned sets of 100,000 reads as input. Output contig sequences of length greater than or equal to the read length are then combined into a single file and further analyzed using the OLC de novo assembler Minimo (Minimo v1.6 release) (Treangen et al. 2011) at default parameters. Contigs are retained if they are >1.75× the length of the original reads. Finally, the full-length unmatched reads, along with the final assembled contigs, are subjected to a protein homology search against the viral protein DB using RAPSearch at an E-value cutoff of 10^{-1}. A user-defined option also allows for a protein homology search against the NCBI nr DB. Retrieved taxonomic information and sequences in FASTA format are appended to the RAPSearch output.

To generate coverage maps, reads classified by SURPI as viral or bacterial are automatically mapped to the most likely reference genome present as follows. For each discrete viral or bacterial genus, assigned NGS reads are directly mapped to all nucleotide reference sequences corresponding to that genus at the species, strain, or substrain level using BLASTn at an E-value cutoff of 10^{-20}. For each genus, a coverage map of the reference sequence with the highest percent coverage is generated, with priority given to reference sequences in the following order: (1) complete genomes; (2) complete sequences; or (3) partial sequences/individual genes.

Detection of clinically relevant pathogens using SURPI

The output of the SURPI pipeline includes a list of all classified reads annotated with their taxonomic assignment; a summary table of read counts stratified by family, genus, species, and accession number (Supplemental Methods); and a series of coverage maps for detected microbial genomes (Supplemental Fig. S2). Coverage maps shown in Figure 5 were edited using Microsoft Excel, as were pie charts derived from the summary tables (Supplemental Tables S6–S21). Sequences corresponding to bacteriophages were grouped together in a single category.

Speed benchmarking for SURPI

End-to-end processing times for the SURPI pipeline (Fig. 7; Supplemental Table S1) were measured using the elapsed wall-clock time and included the following individually timed steps: (1) preprocessing; (2) computational subtraction against the human DB; (3) SNAP alignment to the bacterial DB (*fast* mode); (4) SNAP alignment to the viral nucleotide DB (*fast* mode); (5) SNAP alignment to the complete NCBI nt DB (*comprehensive* mode); (6) de novo contig assembly (*comprehensive* mode); (7) RAPSearch viral protein homology search using translated nucleotide queries (*comprehensive* mode); and (8) overhead time, including file conversion, sequence retrieval, determination of read counts, and generation of summary tables and coverage maps. Processing time trend lines and regression R^2 values were generated using Microsoft Excel.

Data access

The genome sequence of the human parechovirus 1 (HPeV-1) strain described in this study has been deposited in GenBank (http://www.ncbi.nlm.nih.gov/genbank/) as HPeV-1 isolate MX1 (KJ152442). NGS data used for SURPI analysis with potentially identifiable human sequences removed have been submitted to the NCBI Sequence Read Archive (SRA; http://www.ncbi.nlm.nih.gov/sra) under accession number SRP035368. The SURPI software and corresponding Amazon Elastic Compute Cloud (EC2) Amazon Machine Image (AMI) are freely accessible at http://chiulab.ucsf.edu/surpi. The source code for SURPI v1.0 is also available as Supplemental Material.

Competing interest statement

This study was funded in part by an Abbott Viral Discovery Award. C.Y.C. is the director of the UCSF-Abbott Viral Diagnostics and Discovery Center (VDDC). B.S.S. and J.N.F. are employed by a for-profit research organization, Metabiota. K.-C.L. and J.H., Jr. are employed by a for-profit company, Abbott Laboratories.

Acknowledgments

We thank Guixia Yu, Eunice Chen, and Stephanie Yen for processing NGS libraries; William J. Bolosky and Kristal Curtis for their work on SNAP; Andrew Sanderson, Shimona Carvalho, and Geoffrey Hulette for their helpful input on pipeline design; and Eric Delwart for constructive feedback. This work is supported by National Institutes of Health (NIH) grant R01-HL105704 (C.Y.C.), a University of California Discovery Grant (C.Y.C.), the Defense Threat Reduction Agency Cooperative Biological Engagement Program (DTRA-CBEP) (B.S.S., J.N.F., E.L., and C.Y.C.), the United States Agency for International Development (USAID) Emerging Pandemic Threats PREDICT Program (B.S.S., J.N.F., E.L., and C.Y.C.), the AWS in Education Research Grants Program (S.F., N.V., and C.Y.C.), and an Abbott Viral Discovery Award (C.Y.C.). The contents are the responsibility of the authors and do not necessarily reflect the views of USAID or the United States Government.

Author contributions: S.N.N., S.F., N.V., and C.Y.C. developed and benchmarked the pipeline. S.N.N., S.F., N.V., M.Z., T.S., and C.Y.C. developed computational tools for NGS analysis. S.N.N., D.L., E.S., J.B., and A.L.G. generated NGS libraries and validated results using PCR-based assays. S.N.N., S.F., D.L., E.S., J.B., and C.Y.C. ran the pipeline and performed the NGS analysis. B.E., D.W., and S.L.M. provided serum samples from patients with hantavirus pulmonary syndrome. K.-C.L. and J.H., Jr. provided HIV-spiked plasma. G.G., E.L., B.S.S., and J.N.F. provided and coordinated samples from patients with hemorrhagic fever of unknown etiology. J.C. provided samples from patients in Tanzania with febrile illness. M.M. and P.I. contributed stool samples from pediatric diarrheal patients from Mexico. S.M., J.L.D., and C.Y.C. contributed and coordinated samples for clinical testing. S.N.N., S.F., and C.Y.C. wrote the paper.

References

Akobeng AK. 2007. Understanding diagnostic tests 3: receiver operating characteristic curves. *Acta Paediatr* **96:** 644–647.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215:** 403–410.

Barnes GL, Uren E, Stevens KB, Bishop RF. 1998. Etiology of acute gastroenteritis in hospitalized children in Melbourne, Australia, from April 1980 to March 1993. *J Clin Microbiol* **36:** 133–138.

Bhaduri A, Qu K, Lee CS, Ungewickell A, Khavari PA. 2012. Rapid identification of non-human sequences in high-throughput sequencing datasets. *Bioinformatics* **28:** 1174–1175.

Bloch KC, Glaser C. 2007. Diagnostic approaches for patients with suspected encephalitis. *Curr Infect Dis Rep* **9:** 315–322.

Borozan I, Wilson S, Blanchette P, Laflamme P, Watt SN, Krzyzanowski PM, Sircoulomb F, Rottapel R, Branton PE, Ferretti V. 2012. CaPSID: a bioinformatics platform for computational pathogen sequence identification in human genomes and transcriptomes. *BMC Bioinformatics* **13:** 206.

Briese T, Paweska JT, McMullan LK, Hutchison SK, Street C, Palacios G, Khristova ML, Weyer J, Swanepoel R, Egholm M, et al. 2009. Genetic detection and characterization of Lujo virus, a new hemorrhagic fever-associated arenavirus from southern Africa. *PLoS Pathog* **5:** e1000455.

Castellarin M, Warren RL, Freeman JD, Dreolini L, Krzywinski M, Strauss J, Barnes R, Watson P, Allen-Vercoe E, Moore RA, et al. 2012. *Fusobacterium nucleatum* infection is prevalent in human colorectal carcinoma. *Genome Res* **22:** 299–306.

Chen EC, Yagi S, Kelly KR, Mendoza SP, Tarara RP, Canfield DR, Maninger N, Rosenthal A, Spinner A, Bales KL, et al. 2011. Cross-species transmission of a novel adenovirus associated with a fulminant pneumonia outbreak in a new world monkey colony. *PLoS Pathog* **7:** e1002155.

Chiu CY. 2013. Viral pathogen discovery. *Curr Opin Microbiol* **16:** 468–478.

Collins FS, Hamburg MA. 2013. First FDA authorization for next-generation sequencer. *N Engl J Med* **369:** 2369–2371.

Crump JA, Morrissey AB, Nicholson WL, Massung RF, Stoddard RA, Galloway RL, Ooi EE, Maro VP, Saganda W, Kinabo GD, et al. 2013. Etiology of severe non-malaria febrile illness in Northern Tanzania: a prospective cohort study. *PLoS Negl Trop Dis* **7:** e2324.

Delwart EL. 2007. Viral metagenomics. *Rev Med Virol* **17:** 115–131.

Denno DM, Shaikh N, Stapp JR, Qin X, Hutter CM, Hoffman V, Mooney JC, Wood KM, Stevens HJ, Jones R, et al. 2012. Diarrhea etiology in a pediatric emergency department: a case control study. *Clin Infect Dis* **55:** 897–904.

Dimon MT, Wood HM, Rabbitts PH, Arron ST. 2013. IMSA: integrated metagenomic sequence analysis for identification of exogenous reads in a host genomic background. *PLoS ONE* **8:** e64546.

Dunne WM Jr, Westblade LF, Ford B. 2012. Next-generation and whole-genome sequencing in the diagnostic clinical microbiology laboratory. *Eur J Clin Microbiol Infect Dis* **31:** 1719–1726.

Firth C, Lipkin WI. 2013. The genomics of emerging pathogens. *Annu Rev Genomics Hum Genet* **14:** 281–300.

Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S, et al. 2002. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419:** 498–511.

Gargis AS, Kalman L, Berry MW, Bick DP, Dimmock DP, Hambuch T, Lu F, Lyon E, Voelkerding KV, Zehnbauer BA, et al. 2012. Assuring the quality of next-generation sequencing in clinical laboratory practice. *Nat Biotechnol* **30:** 1033–1036.

Grard G, Fair JN, Lee D, Slikas E, Steffen I, Muyembe JJ, Sittler T, Veeraraghavan N, Ruby JG, Wang C, et al. 2012. A novel rhabdovirus associated with acute hemorrhagic fever in central Africa. *PLoS Pathog* **8:** e1002924.

Gremme G, Steinbiss S, Kurtz S. 2013. GenomeTools: a comprehensive software library for efficient processing of structured genome annotations. *IEEE/ACM Trans Comput Biol Bioinformatics* **10:** 645–656.

Greninger AL, Chen EC, Sittler T, Scheinerman A, Roubinian N, Yu G, Kim E, Pillai DR, Guyard C, Mazzulli T, et al. 2010. A metagenomic analysis of pandemic influenza A (2009 H1N1) infection in patients from North America. *PLoS ONE* **5:** e13381.

Knope K, Whelan P, Smith D, Johansen C, Moran R, Doggett S, Sly A, Hobby M, Kurucz N, Wright P, et al. 2013. Arboviral diseases and malaria in Australia, 2010-11: annual report of the National Arbovirus and Malaria Advisory Committee. *Commun Dis Intell Q Rep* **37:** E1–E20.

Kollef KE, Schramm GE, Wills AR, Reichley RM, Micek ST, Kollef MH. 2008. Predictors of 30-day mortality and hospital costs in patients with ventilator-associated pneumonia attributed to potentially antibiotic-resistant gram-negative bacteria. *Chest* **134:** 281–287.

Kostic AD, Ojesina AI, Pedamallu CS, Jung J, Verhaak RG, Getz G, Meyerson M. 2011. PathSeq: software to identify or discover microbes by deep sequencing of human tissue. *Nat Biotechnol* **29:** 393–396.

Kostic AD, Gevers D, Pedamallu CS, Michaud M, Duke F, Earl AM, Ojesina AI, Jung J, Bass AJ, Tabernero J, et al. 2012. Genomic analysis identifies association of *Fusobacterium* with colorectal carcinoma. *Genome Res* **22:** 292–298.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9:** 357–359.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25:** 1754–1760.

Loman NJ, Constantinidou C, Chan JZ, Halachev M, Sergeant M, Penn CW, Robinson ER, Pallen MJ. 2012. High-throughput bacterial genome

sequencing: an embarrassment of choice, a world of opportunity. *Nat Rev Microbiol* **10:** 599–606.
Louie JK, Hacker JK, Gonzales R, Mark J, Maselli JH, Yagi S, Drew WL. 2005. Characterization of viral agents causing acute respiratory infection in a San Francisco University Medical Center Clinic during the influenza season. *Clin Infect Dis* **41:** 822–828.
MacConaill L, Meyerson M. 2008. Adding pathogens by genomic subtraction. *Nat Genet* **40:** 380–382.
Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetjournal* **17:** 1.
Morse SS, Mazet JA, Woolhouse M, Parrish CR, Carroll D, Karesh WB, Zambrana-Torrelio C, Lipkin WI, Daszak P. 2012. Prediction and prevention of the next pandemic zoonosis. *Lancet* **380:** 1956–1965.
Naeem R, Rashid M, Pain A. 2013. READSCAN: a fast and scalable pathogen discovery program with accurate genome relative abundance estimation. *Bioinformatics* **29:** 391–392.
Niu B, Zhu Z, Fu L, Wu S, Li W. 2011. FR-HIT, a very fast program to recruit metagenomic reads to homologous reference genomes. *Bioinformatics* **27:** 1704–1705.
Nunez JJ, Fritz CL, Knust B, Buttke D, Enge B, Novak MG, Kramer V, Osadebe L, Messenger S, Albarino CG, et al. 2014. Hantavirus infections among overnight visitors to Yosemite National Park, California, USA, 2012. *Emerg Infect Dis* **20:** 386–393.
Pruitt KD, Tatusova T, Maglott DR. 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **35:** D61–D65.
Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y. 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* **13:** 341.
Schmieder R, Edwards R. 2011. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27:** 863–864.
Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I. 2009. ABySS: a parallel assembler for short read sequence data. *Genome Res* **19:** 1117–1123.
Sommer DD, Delcher AL, Salzberg SL, Pop M. 2007. Minimus: a fast, lightweight genome assembler. *BMC Bioinformatics* **8:** 64.
Swei A, Russell BJ, Naccache SN, Kabre B, Veeraraghavan N, Pilgard MA, Johnson BJ, Chiu CY. 2013. The genome sequence of Lone Star virus, a highly divergent bunyavirus found in the *Amblyomma americanum* tick. *PLoS ONE* **8:** e62083.
Tong S, Li Y, Rivailler P, Conrardy C, Castillo DA, Chen LM, Recuenco S, Ellison JA, Davis CT, York IA, et al. 2012. A distinct lineage of influenza A virus from bats. *Proc Natl Acad Sci* **109:** 4269–4274.
Treangen TJ, Sommer DD, Angly FE, Koren S, Pop M. 2011. Next generation sequence assembly with AMOS. *Curr Protoc Bioinformatics* **33:** 11.8.1–11.8.18.
van Gageldonk-Lafeber AB, Heijnen ML, Bartelds AI, Peters MF, van der Plas SM, Wilbrink B. 2005. A case-control study of acute respiratory tract infection in general practice patients in The Netherlands. *Clin Infect Dis* **41:** 490–497.
Wang Q, Jia P, Zhao Z. 2013. VirusFinder: software for efficient and accurate detection of viruses and their integration sites in host genomes through next generation sequencing data. *PLoS ONE* **8:** e64465.
Ward KN, Kalima P, MacLeod KM, Riordan T. 2002. Neuroinvasion during delayed primary HHV-7 infection in an immunocompetent adult with encephalitis and flaccid paralysis. *J Med Virol* **67:** 538–541.
Wilson MR, Naccache SN, Samayoa E, Biagtan M, Bashir H, Yu G, Salamat SM, Somasekar S, Federman S, Miller S, et al. 2014. Actionable diagnosis of neuroleptospirosis by next-generation sequencing. *N Engl J Med* doi: 10.1056/NEJMoa.1401268.
Wylie KM, Mihindukulasuriya KA, Sodergren E, Weinstock GM, Storch GA. 2012. Sequence analysis of the human virome in febrile and afebrile children. *PLoS ONE* **7:** e27735.
Xu B, Liu L, Huang X, Ma H, Zhang Y, Du Y, Wang P, Tang X, Wang H, Kang K, et al. 2011. Metagenomic analysis of fever, thrombocytopenia and leukopenia syndrome (FTLS) in Henan Province, China: discovery of a new bunyavirus. *PLoS Pathog* **7:** e1002369.
Yu G, Greninger AL, Isa P, Phan TG, Martinez MA, de la Luz Sanchez M, Contreras JF, Santos-Preciado JI, Parsonnet J, Miller S, et al. 2012. Discovery of a novel polyomavirus in acute diarrheal samples from children. *PLoS ONE* **7:** e49449.
Zaharia M, Bolosky WJ, Curtis K, Fox A, Patterson D, Shenker S, Stoica I, Karp RM, Sittler T. 2011. Faster and more accurate sequence alignment with SNAP. *arXiv* 1111.5572.
Zaki AM, van Boheemen S, Bestebroer TM, Osterhaus AD, Fouchier RA. 2012. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. *N Engl J Med* **367:** 1814–1820.
Zhao Y, Tang H, Ye Y. 2012. RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. *Bioinformatics* **28:** 125–126.
Zhao G, Krishnamurthy S, Cai Z, Popov VL, Travassos da Rosa AP, Guzman H, Cao S, Virgin HW, Tesh RB, Wang D. 2013. Identification of novel viruses using VirusHunter–an automated data analysis pipeline. *PLoS ONE* **8:** e78470.
Zweig MH, Campbell G. 1993. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin Chem* **39:** 561–577.

Received December 31, 2013; accepted in revised form March 26, 2014.

A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples

Samia N. Naccache, Scot Federman, Narayanan Veeraraghavan, et al.

Genome Res. 2014 24: 1180-1192 originally published online June 4, 2014
Access the most recent version at doi:10.1101/gr.171934.113