Comprehensive human virus screening using high-throughput sequencing with a user-friendly representation of bioinformatics analysis: a pilot study

PETTY II, Thomas John, et al.

Abstract

High-throughput sequencing (HTS) provides the means to analyze clinical specimens in unprecedented molecular detail. While this technology has been successfully applied to virus discovery and other related areas of research, HTS methodology has yet to be exploited for use in a clinical setting for routine diagnostics. Here, a bioinformatics pipeline (ezVIR) was designed to process HTS data from any of the standard platforms and to evaluate the entire spectrum of known human viruses at once, providing results that are easy to interpret and customizable. The pipeline works by identifying the most likely viruses present in the specimen given the sequencing data. Additionally, ezVIR can generate optional reports for strain typing, can create genome coverage histograms, and can perform cross-contamination analysis for specimens prepared in series. In this pilot study, the pipeline was challenged using HTS data from 20 clinical specimens representative of those most often collected and analyzed in daily practice. The specimens (5 cerebrospinal fluid, 7 bronchoalveolar lavage fluid, 5 plasma, 2 serum, and 1 nasopharyngeal [...]

Reference


DOI : 10.1128/JCM.01389-14
PMID : 25009045

Available at:
http://archive-ouverte.unige.ch/unige:74559

Disclaimer: layout of this document may differ from the published version.
Comprehensive Human Virus Screening Using High-Throughput Sequencing with a User-Friendly Representation of Bioinformatics Analysis: a Pilot Study

Tom J. Petty, a,b Samuel Cordey, c,d Ismael Padioleau, a,b* Mylène Docquier, e Lara Turin, d Olivier Preynat-Seauve, f Evgeny M. Zdobnov, a,b,g Laurent Kaiser c,d

Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, Switzerland; Swiss Institute of Bioinformatics, Geneva, Switzerland; Division of Infectious Diseases, Laboratory of Virology and Division of Laboratory Medicine, University Hospitals of Geneva, Geneva, Switzerland; Department of Medicine, University of Geneva Medical School, Geneva, Switzerland; University of Geneva Medical School and Genomics Platform, Geneva, Switzerland; Laboratory of Immunohematology, Hematology Unit, Department of Genetic and Laboratory Medicine, University Hospitals of Geneva, Geneva, University of Geneva, Geneva, Switzerland; Laboratory of Immunohematology, Hematology Unit, Department of Genetic and Laboratory Medicine, University Hospitals of Geneva, Geneva, University of Geneva, Geneva, Switzerland; Imperial College London, South Kensington Campus, London, United Kingdom

High-throughput sequencing (HTS) provides the means to analyze clinical specimens in unprecedented molecular detail. While this technology has been successfully applied to virus discovery and other related areas of research, HTS methodology has yet to be exploited for use in a clinical setting for routine diagnostics. Here, a bioinformatics pipeline (ezVIR) was designed to process HTS data from any of the standard platforms and to evaluate the entire spectrum of known human viruses at once, providing results that are easy to interpret and customizable. The pipeline works by identifying the most likely viruses present in the specimens given the sequencing data. Additionally, ezVIR can generate optional reports for strain typing, can create genome coverage histograms, and can perform cross-contamination analysis for specimens prepared in series. In this pilot study, the pipeline was challenged using HTS data from 20 clinical specimens representative of those most often collected and analyzed in daily practice. The specimens (5 cerebrospinal fluid, 7 bronchoalveolar lavage fluid, 5 plasma, 2 serum, and 1 nasopharyngeal aspirate) were originally found to be positive for a diverse range of DNA or RNA viruses by routine molecular diagnostics. The ezVIR pipeline correctly identified 14 of 14 specimens containing viruses with genomes of <40,000 bp, and 4 of 6 specimens positive for large-genome viruses. Although further validation is needed to evaluate sensitivity and to define detection cutoffs, results obtained in this pilot study indicate that the overall detection success rate, coupled with the ease of interpreting the analysis reports, makes it worth considering using HTS for clinical diagnostics.

Over the last decade, high-throughput sequencing (HTS) has provided unprecedented opportunities for advancement in the field of virology (1–3). To date, HTS has been widely used for microbiome analysis (4–7), whole-genome sequencing (8–12), quasispecies population analysis (13–16), and the discovery of novel viruses (17–21). As a result, various methods and web-based services, such as VIDISCA-454 (22), Pathoscope (23), MetaVir (24), VirusHunter (25), or VirusFinder (26), have emerged. However, each is specific to a particular HTS platform, and interpretation of results remains convoluted to non-experts in bioinformatics. Yet it is reasonable to consider that this technology will soon be adopted for use in routine clinical diagnostics, as HTS has the potential to improve standard diagnostics in many ways, such as providing the means to identify unexpected pathogens, increasing assay sensitivity, and detecting viruses for which no assay exists. Even commercial HTS-based virus typing assays are beginning to emerge, such as the PathAmp FluA reagent (Life Technologies). An additional motivating factor for using HTS is that infectious disease specialists must often handle cases where viral origin is highly suspected, but where all microbiological test results are negative or inconclusive. HTS can provide the means for an alternative diagnostic tool to analyze such specific, potentially life-threatening cases.

To this end, the commonly used Roche-454 and Illumina HTS platforms were recently evaluated for their capacity to detect viruses either using artificially spiked specimens (27) or with specific preselected clinical specimens (22, 28). However, in the latter studies, the analysis and interpretation were focused on the capacity of HTS to detect the known target(s). Consequently, there was little evaluation of the capacity of these methodologies for clinical diagnostics, a situation where “background” sequences play an important role in how results are interpreted. For example, clinical specimens often contain common circulating viruses, such as Epstein-Barr virus (EBV), human herpesvirus 6 (HHV-6), and torque teno virus (TTV), whose presence and quantity vary with specimen type (e.g., blood versus cerebrospinal fluid) and can potentially mask the signal of the causative pathogen.

Therefore, there is a need to benchmark the ability of HTS to...
provide clear and unbiased viral diagnostic results using a set of diverse, clinically relevant positive specimens. The current lack of diagnostics-oriented HTS investigations in the literature is likely due to the difficulty in displaying and interpreting the complicated data. Often, such results are incomprehensible to non-specialists in bioinformatics. The current lack of sufficient amounts of sequence data per specimen. In this pilot study, we present ezVIR as a proof of principle for using HTS for clinical diagnostics. We aimed to demonstrate that HTS can be effectively used in a clinical setting to help non-specialists in bioinformatics make more informed decisions, as all detection results are easy to understand. This bioinformatics pipeline was developed using positive clinical specimens and is designed to processes HTS data and provide easy-to-interpret results independently of the HTS platform used. Given the elevated sensitivity of HTS technology, we also designed a tool to identify cross-contamination in situations where specimens were prepared simultaneously. The results are reported in two phases to enable rapid identification and subsequent typing of the identified viruses, including the ability to customize the reports.

**MATERIALS AND METHODS**

**Clinical specimen selection.** Cerebrospinal fluid (CSF), bronchoalveolar lavage (BAL) fluid, serum, plasma, and nasopharyngeal aspirate (NPA) analyzed in this pilot study were selected from specimens submitted between January and September 2013 to the Laboratory of Virology from the University Hospitals of Geneva, Switzerland. The patient’s ages ranged from 10 months to 88 years (median, 42 years old). Specimens were randomly selected with the criteria of obtaining one representative of different clinically relevant virus species and genome properties per specimen type, each with an average viral load determined by semiquantitative (e.g., threshold cycle \( C_{T} \)) and quantitative (quantitative PCR [qPCR]) molecular testing (Table 1). Specimens were identified to be positive for either

---

**TABLE 1 Specimen information and HTS summary statistics**

<table>
<thead>
<tr>
<th>Sample</th>
<th>Virus*</th>
<th>Genome</th>
<th>Specimen type</th>
<th>Routine screening value</th>
<th>HTS sequencing</th>
<th>Identification by ezVIR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>DNA</td>
<td>RNA</td>
<td>CSF BAL NPA Plasma Serum</td>
<td>Total no. of reads</td>
<td>% human reads§</td>
</tr>
<tr>
<td>01</td>
<td>HSV-1</td>
<td>X</td>
<td>X</td>
<td>C(_{T}), 28.3</td>
<td>30,393,082</td>
<td>2.76</td>
</tr>
<tr>
<td>02</td>
<td>JCV</td>
<td>X</td>
<td>X</td>
<td>2.3E6 copies/ml</td>
<td>26,716,996</td>
<td>4.00</td>
</tr>
<tr>
<td>03</td>
<td>VZV</td>
<td>X</td>
<td>X</td>
<td>C(_{T}), 30.2</td>
<td>169,259,146</td>
<td>26.89</td>
</tr>
<tr>
<td>04</td>
<td>CMY</td>
<td>X</td>
<td>X</td>
<td>139 IU/ml‡</td>
<td>171,327,278</td>
<td>26.21</td>
</tr>
<tr>
<td>05</td>
<td>BKV</td>
<td>X</td>
<td>X</td>
<td>2.5E4 copies/ml</td>
<td>135,514,616</td>
<td>6.50</td>
</tr>
<tr>
<td>06</td>
<td>EBV</td>
<td>X</td>
<td>X</td>
<td>1.3E3 copies/ml</td>
<td>101,791,438</td>
<td>1.41</td>
</tr>
<tr>
<td>07</td>
<td>CMV</td>
<td>X</td>
<td>X</td>
<td>736 IU/ml</td>
<td>238,166,770</td>
<td>4.21</td>
</tr>
<tr>
<td>08</td>
<td>HSV-2</td>
<td>X</td>
<td>X</td>
<td>C(_{T}), 30</td>
<td>33,127,084</td>
<td>0.65</td>
</tr>
<tr>
<td>09</td>
<td>ParvoB19</td>
<td>X</td>
<td>X</td>
<td>1.2E5 IU/ml</td>
<td>88,326,612</td>
<td>14.31</td>
</tr>
<tr>
<td>10</td>
<td>EchoV6</td>
<td>X</td>
<td>X</td>
<td>C(_{T}), 27.6</td>
<td>19,467,996</td>
<td>6.11</td>
</tr>
<tr>
<td>11</td>
<td>HRV-73</td>
<td>X</td>
<td>X</td>
<td>C(_{T}), 21.6</td>
<td>20,802,852</td>
<td>10.71</td>
</tr>
<tr>
<td>12</td>
<td>EchoV30</td>
<td>X</td>
<td>X</td>
<td>C(_{T}), 27.7</td>
<td>69,377,802</td>
<td>12.59</td>
</tr>
<tr>
<td>13</td>
<td>MeV B3</td>
<td>X</td>
<td>X</td>
<td>C(_{T}), 20.5</td>
<td>61,667,042</td>
<td>12.84</td>
</tr>
<tr>
<td>14</td>
<td>hMPV</td>
<td>X</td>
<td>X</td>
<td>C(_{T}), 18</td>
<td>50,671,732</td>
<td>9.62</td>
</tr>
<tr>
<td>15</td>
<td>hPIV-3</td>
<td>X</td>
<td>X</td>
<td>C(_{T}), 24.8</td>
<td>42,329,200</td>
<td>11.33</td>
</tr>
<tr>
<td>16</td>
<td>HRV-66</td>
<td>X</td>
<td>X</td>
<td>C(_{T}), 17.8</td>
<td>54,192,578</td>
<td>9.54</td>
</tr>
<tr>
<td>17</td>
<td>IAV</td>
<td>X</td>
<td>X</td>
<td>C(_{T}), 31</td>
<td>65,744,586</td>
<td>19.04</td>
</tr>
<tr>
<td>18</td>
<td>HIV-1</td>
<td>X</td>
<td>X</td>
<td>3.5E5 copies/ml</td>
<td>49,268,460</td>
<td>10.88</td>
</tr>
<tr>
<td>19</td>
<td>HCV</td>
<td>X</td>
<td>X</td>
<td>1.3E6 copies/ml</td>
<td>81,148,650</td>
<td>7.73</td>
</tr>
<tr>
<td>20</td>
<td>PeV</td>
<td>X</td>
<td>X</td>
<td>C(_{T}), 29.9</td>
<td>145,792,742</td>
<td>83.63</td>
</tr>
<tr>
<td></td>
<td>PIV-1</td>
<td>X</td>
<td>X</td>
<td>C(_{T}), 35.5</td>
<td>8,168,450</td>
<td>7.73</td>
</tr>
<tr>
<td></td>
<td>PIV-3</td>
<td>X</td>
<td>X</td>
<td>C(_{T}), 30.9</td>
<td>65,744,586</td>
<td>19.04</td>
</tr>
<tr>
<td></td>
<td>PIV-4</td>
<td>X</td>
<td>X</td>
<td>C(_{T}), 31.4</td>
<td>81,148,650</td>
<td>7.73</td>
</tr>
<tr>
<td></td>
<td>MeV</td>
<td>X</td>
<td></td>
<td>C(_{T}), 29.4</td>
<td>145,792,742</td>
<td>83.63</td>
</tr>
</tbody>
</table>

* CSF, cerebrospinal fluid; BAL, bronchoalveolar lavage fluid; NPA, nasopharyngeal aspirate; C\(_{T}\), threshold cycle; IU, international units; §, properly paired reads; ‡, at lower limit of detection; ✓, detected; —, inconclusive.

* HSV-1 and -2, herpes simplex virus 1 and 2, respectively; JCV, JC virus; VZV, varicella-zoster virus; CMY, cytomegalovirus; BKV, BK virus; EBV, Epstein-Barr virus; ParvoB19, parvovirus B19; EchoV6 and EchoV30, echovirus 6 and 30, respectively; HRV-73 and -66, human rhinovirus 73 and 66, respectively; MeV B3 and MeV Edmonston, measles virus genotype B3 and Edmonston strain, respectively; hMPV, human metapneumovirus; hPIV-3, human parainfluenza type 3; IAV, influenza A virus; HIV-1, human immunodeficiency virus type 1; HCV, hepatitis C virus; PeV, parechovirus; PIV-1, -3, and -4, parainfluenza type 1, 3, and 4, respectively.

Petty et al.
human rhinovirus (HRV), echovirus (EchoV; an enterovirus), parechovirus (PeV), influenza A virus (IAV), measles virus (MeV), human metapneumovirus (hMPV), human parainfluenza type 1 (hPIV-1), hPIV-3, hPIV-4, human immunodeficiency virus type 1 (HIV-1), hepatitis C virus (HCV), parvovirus B19 (ParvoB19), herpes simplex virus 1 (HSV-1), HSV-2, varicella-zoster virus (VZV; also known as human herpesvirus 3), cytomegalovirus (CMV; human herpesvirus 5), Epstein-Barr virus (EBV; human herpesvirus 4), BK virus (BKV), or JC virus (JCV) by in-house assays or commercial real-time PCR (RT-PCR). After routine screening, each specimen was immediately stored at −20°C (blood specimens) or −80°C (CSF, BAL fluid, NPA). The ethics committee of the Geneva University Hospitals approved this study and determined that no informed consent was required.

Viral nucleic acid extraction. For each specimen analyzed, 110 μl was centrifuged at 10,000 × g for 10 min. One hundred microliters of cell-free supernatant was collected and treated with 20 U of Turbo DNase (Ambion, Rotkreuz, Switzerland) to degrade non-particle-protected DNA. The RNA pellet was resuspended in 20 μl of RNase-free water (Promega, Dübendorf, Switzerland). DNA virus genome extraction was performed with Trizol according to the manufacturer’s instructions (Invitrogen, Carlsbad, CA, USA). DNA virus genome extraction was performed with a NucliSens easyMAG magnetic bead system (bioMérieux, Geneva, Switzerland) according to the manufacturer’s instructions, using an elution volume of 25 μl. Subsequent double-stranded DNA synthesis was performed as described by De Vries et al. (22, 28) with some modifications. A 120-μl mixture containing 5 U of a 3’-exo−Klenow fragment (New England BioLabs, Ipswich, MA, USA), 3 μg of random hexamers (Invitrogen), 0.4× Escherichia coli ligation buffer (Invitrogen), 1.8 mM MgCl2, 0.75 mM dithiothreitol (DTT), 0.3 mM dNTPs, and 16 μg/ml RNase A (Sigma-Aldrich, Buchs, Switzerland) was added to 20 μl of extracted DNA, incubated 1.5 h at 37°C, and subjected to a phenol-chloroform extraction and ethanol precipitation. The DNA pellet was resuspended in 20 μl of RNase-free water (Promega).

High-throughput DNA sequencing (DNA-seq) library preparation. Nine specimens were prepared. The volume of the specimens was reduced to 5 μl to measure the concentration. For 8 of the 9 specimens, the amount of starting material was 1 ng (representing approximately one-third of the total amount of material). The totality of specimen 09 was used, since the concentration was below the limit of detection. Libraries were prepared using the Illumina Nextera XT protocol (12 PCR cycles). Library concentrations were measured with a Q-bit (Life Technologies, Carlsbad, CA, USA). The most concentrated library (specimen 03) was detectable by a 2200 TapeStation (Agilent, Santa Clara, CA, USA). Fragments of 150 to 450 bp were obtained.

High-throughput RNA sequencing (RNA-seq) library preparation. Eleven specimens were prepared. The starting amount was unknown (below the limit of detection) for 10 of the 11 specimens. For specimen 17, the starting material was 80 ng. The RNA was removed using a Ribo-Zero Gold kit (epicenter, Madison, WI, USA) according to the manufacturer’s protocol. RNA-depleted specimens were purified on Zymo columns. For specimen 20, a poly(A) depletion using a TruSeq Stranded mRNA kit (Illumina, San Diego, CA, USA) was performed after removal of RNA. Libraries were prepared with the low-throughput TruSeq total RNA preparation protocol from Illumina (San Diego, CA, US) using 15 PCR cycles. Library concentrations were measured with a Q-bit. Size distribution of fragments was estimated with a 2200 TapeStation. Fragments of 200 to 450 bp were obtained.

High-throughput sequencing. All specimens were sequenced (paired end [PE]) using the 100-bp protocol with indexing on a HiSeq 2500 (Il- luminia) sequencer in pools of two specimens per lane. RNA-seq libraries were loaded at 8 pM, DNA-seq libraries were loaded at 20 pM or lower for low-concentrated libraries (specimen 01, specimen 02, and specimen 08 were loaded at 10, 13, and 16 pM, respectively). The specimen 03 library size was used to calculate molarity.

Virus database. Complete mammalian and avian virus genome sequences were collected from EMBL, ViralZone, and NCBI databases, as no single comprehensive collection of full-length virus genome sequences currently exists (for more detail, see Table S1 in the supplemental material). Briefly, all genomes listed in both EMBL virus (http://www.ebi.ac.uk/genomes/virus.html) and ViralZone (http://viralzone.expasy.org/all_by_species/678.html) collections were merged into one database. Complete virus sequences from any missing virus families were then downloaded from NCBI and combined with the EMBL and ViralZone sequences. All duplicates and any unverified genomes were removed. Furthermore, any genomes labeled as “recombinant” or “clone” were carefully inspected and conserved only if pertinent.

Virus genome bin size selection. A histogram of genome lengths for all genomes in the ezVIR database was generated in order to determine optimal bin size values for the data point dot circumference (genome size) groups shown on the ezVIR plots (see Fig. S1 in the supplemental material) (Fig. 1C). The default bin cutoff values are 4,000, 20,000, and 40,000 bp; however, these cutoff values can be customized by the user in order to modify the appearance of viruses on result plots, depending on the aim of the particular study or application.

Bioinformatics pipeline. The specimen libraries were bar-coded and sequenced in a multiplexed reaction, and the resulting PE reads (100 nucleotides each) were demultiplexed (libraries were separated by their index). To remove human sequences, reads are first mapped (Bowtie2) (30) to the human genome (NCBI GRCh37). In virus identification phase 1, the remaining nonhuman reads are mapped (Bowtie2) to a comprehensive, manually curated database containing 11,018 complete virus genomes (see Table S1 and Fig. S1 in the supplemental material). To increase the sensitivity of detection, all mate-pair mappings, for all reads and for every genome, are retained. By exposing each genome to all reads during this stage, the pipeline is able to determine the most likely viruses (as the best-mapped genome for each virus species [genus for herpesviruses]) given the sequence data. After the mapping stages, the pipeline computes genome detection metrics (defined below), summarizes data, and generates reports at two levels of sensitivity: phase 1, general virus identification (positive targets representing the strongest signal from each virus species [genus for herpesviruses]) detected; and phase 2, targeted strain detection and genome coverage statistics (Fig. 1A).

After mapping to all genomes in the database, three metrics are calculated for every detected virus genome: percent genome coverage, maximum depth of coverage (using a sliding window of 50 bp), and total covered length in bp (Fig. 1B). The percent genome coverage is calculated as the total length of all regions along the genome that are covered by at least one read, divided by genome length. This serves as an intrinsic normalization, as the lengths of virus genomes vary by more than 100 orders of magnitude. The maximum depth of coverage reflects the relative “signal strength” of a particular virus and is calculated as the maximum of the average number of reads in a sliding window (default size of 50 bp) along the genome. The sliding window provides the means to highlight viruses with slightly more genome coverage in cases where multiple viruses have the same upper limit of mapped reads (for more detail, see Fig. S2 in the supplemental material). The total covered length indicates the total number of nucleotides detected for each virus genome and is represented on the reports) by colored dots of varied circumference (larger circumferences correspond to more nucleotides covered). Phase 1 plots display the best-scoring representative (highest percent coverage and greatest maximum depth of coverage) for each detected virus species. For each virus identified in phase 1, phase 2 provides genome coverage information on the level of strains, genotype, serotype, or lineage, depending on the virus identified. For this pilot study, while all specimens were mapped to the complete database (and therefore retained potential mappings to nonhuman viruses), only known human viruses are shown in the reports. Aside from mapping, all data processing, analysis, statistics calculations, and reporting are coded in, and performed with, BEDtools (31), R (32), python (http://www.python.org), and Linux (bash) shell scripts. Regarding
data storage, the largest files are the initial raw sequencing results (“fastq” files) in the range of 4 to 20 gigabytes (GB) per specimen. All ezVIR analysis files (including all mapping results) are significantly smaller (on the order of megabytes [MB]) and can easily be stored on desktop systems. The code for this pilot version of the ezVIR pipeline and supporting documentation is available at http://cegg.unige.ch/ezvir.

CCA. Genome mapping results for all specimens are compared in a pairwise manner for all possible pair permutations (specimen 01 versus specimen 02, specimen 01 versus specimen 03, specimen 02 versus specimen 03, etc.). Per pair, any virus genomes that are detected in both specimens are stored as an intersect set with corresponding ezVIR detection metrics (percent genome coverage, maximum depth of coverage, total covered length, and genome length, as explained above) per genome. These intersect sets can be queried in phase 2 on a per-virus-species basis. To perform the cross-contamination analysis (CCA), once any detected virus appearing on the report plot is selected, the cross-contamination module will create a bar plot displaying the “signal” (the log10 maximum coverage depth) for that particular virus genome in all specimens (Fig. 2C) (see Fig. S4 in the supplemental material). The CCA plots serve to guide interpretation of ezVIR analysis results and help determine if a detected virus was present in the original specimen or could be a contaminant from a neighboring specimen or the laboratory environment.

Workflow. Viral nucleic acids are extracted directly from clinical specimens without particle enrichment steps and then processed as described above for HTS (Fig. 1A). The input to ezVIR is the sequence data (“fastq” files, the output from all standard HTS platforms), and the default output is the phase 1 report. Based on the viruses identified in this phase, the user can then ask for more detailed information (in phase 2) about each particular virus, including read coverage histograms, strain typing, and cross-contamination reports.

How to interpret reports. The reports are designed to allow rapid and intuitive comparison of all viruses detected regardless of large differences in genome lengths. Only the best-mapped (in terms of genome coverage and maximum depth of coverage) genomes from each virus species (genus for herpesviruses) are presented in the initial phase 1 reports (Fig. 1). Genus- or strain-specific information can be viewed in phase 2 reports. For each detected virus, a dot appears on the plot, reflecting the percent genome coverage on the x-axis (how much of the genome was detected) and the signal strength on the y-axis (the most reads mapped calculated as the maximum depth of coverage). Ideally, the causative and/or most prominent virus in a specimen will appear in the upper right corner, representing 100% genome coverage and the strongest signal (y-axis) relative to those of other potential viruses in the same specimen. Generally, when 100% genome coverage is not observed, the best candidate(s) is represented by the dot corresponding to the clinically relevant virus species or genus that combines the highest percent genome coverage and maximum depth of coverage. Furthermore, the metrics corresponding to the total covered length indicated in phase 1 reports (colored dots of varied circumference) also help to highlight the viruses of interest in such cases. The y-axis scale is dynamic and is automatically adjusted according to the virus with the strongest signal in each specimen (Fig. 1B). Although we do not define a lower limit of detection (cutoff), any virus appearing with a low y-axis value (based on our observations here, a value of <10) must be considered with caution (refer to “Dealing with cross-contamination” below).

The data points (colored dots) that appear on the reports are intended to provide multiple useful measurements in the same location (Fig. 1C, Data point information). The color of the inner dot and label name corresponds to the virus family (key on right side of reports). The outer gray ring represents the size class (genome length) of the virus. The size of the inner dot indicates approximately how much of that genome was detected (in terms of mapped nucleotides). The different dot sizes help to compare viruses of vastly different sizes on the same plot. Additionally, a table in the form of a comma-separated text file (for use with standard data display software, such as Microsoft Excel) that contains all values for all detected viruses accompanies each graphical report (Fig. 3).

Cost and practicality. In this pilot study, each specimen was sequenced (Illumina HiSeq platform) using standard technology with a standard paired-end protocol. The cost of sequencing (including library preparation) was approximately $1,500 per paired-end run. While the first part of the pilot ezVIR pipeline (mapping to the human genome and then to all virus genomes) (Fig. 1A) took ~4 days per specimen using Bowtie2 (30) on a multicore computer (~100 central processing units [CPUs]), the use of alternate alignment software (e.g., SNAP) (33) can reduce the analysis time to less than 1 day. The speed of the mapping stage depends on the mapping software used, the number of computing cores, and the number of nonhuman reads. After mapping, all report generation and the phase 2 analysis can be performed on a desktop or laptop computer in a matter of seconds to minutes.

RESULTS

Bioinformatics pipeline. With the goal of identifying viruses (those for which genome sequences exist) in clinical specimens, our bioinformatics tools are designed to remove human “background” sequences, identify virus genus or species, subsequently identify particular strains, and report findings in a comprehensible manner (Fig. 1, 2, and 3) (see Fig. S3 in the supplemental material). The pipeline generates reports in two stages in order to simplify interpretation without compromising the presentation of significant microbiological findings. The bioinformatics tools presented here have been developed to enable better discrimination of the “true-positive” human viral sequences within the “noise” of multiple background sequences. The main steps involve (i) mapping the HTS data to the human genome, (ii) mapping all nonhuman reads to a comprehensive database of virus genomes, and (iii) computing mapping metrics (see Materials and Methods) and then organizing and summarizing the results into user-friendly and comprehensible graphical reports (Fig. 1A).

To design and subsequently challenge the pipeline, 20 clinical specimens (5 CSF, 7 BAL fluid, 1 NPA, 5 plasma, and 2 serum specimens) found to be positive for either DNA or RNA viruses by routine molecular diagnostics were analyzed (Table 1). After mapping the HTS data from each specimen to the human genome, the number of human reads per specimen was found to vary (from 0.65% to ~84%) according to the clinical specimen. For each specimen, all nonhuman reads were mapped to our database,

FIG 1 ezVIR pipeline, metrics, and reporting. (A) ezVIR pipeline overview. DB, database. (B) Three metrics were reported for each detected virus. Percent genome coverage reports all regions of the genome (blue) covered by at least one read (purple) divided by genome length. Maximum (max) depth of coverage refers to the average number of reads covering the genome in a 50-bp sliding window; the window slides along the genome and the maximum value is reported. Total covered length is the total number of bases detected for the genome. Simple equations demonstrate how each is calculated (results are in green). (C) ezVIR reporting features, including the type of data contained in each report, analysis options, data point information, and clinically relevant virus family colors. ID, identification number. (D) Examples of phase 1 reports (using specimens 01 and 16, which were found to be positive for herpes simplex virus 1 and HRV-66 by traditional routine clinical diagnostics, respectively). Plots depict six metrics per identified virus: (i) virus type, (ii) virus family (color of dot and label), (iii) percent genome coverage, (iv) maximum coverage depth, (v) total covered length in base pairs (represented by the area of the colored dot), (vi) genome size group (gray outer ring).
A  example phase-2 report: HRV typing

B  example co-infection: Parechovirus, PIV-1, -3, -4, and MeV in specimen-20

C  example CCA: HRV in specimen-17

* specimen shown positive for HRV
metrics calculated, data analyzed, and results presented in two phases using the ezVIR tools described below.

**Analysis of positive clinical specimens.** Eighteen of the 20 specimens used to assess the robustness of the pipeline presented here were correctly analyzed: 16 specimens in phase 1, and 2 specimens in phase 2 (specimen 12 and specimen 17). Results were inconclusive at the genus level for the remaining 2 specimens (specimen 03 and specimen 04) (Table 1) (see Fig. S3 in the supplemental material). Of note, as the specimens studied here represent a broad range of those often found in routine clinical situations, we also gained valuable insight into areas for improvement and the potential limitations (described in the following sections) of using HTS in a clinical setting.

The pipeline generates reports in two phases (Fig. 1A). The phase 1 report serves as the default representation that indicates the strongest signal from each detected virus species (Fig. 1D). To reduce background signals (viruses with very low genome coverage) and improve the interpretation of the results, users have the option to define a threshold value for “percent genome coverage” (for example, the threshold can be set to display only those viruses with more than 5% genome coverage). Since the y-axis scale is dynamic, this option may be useful to better differentiate partially overlapping dots (including the associated labels). The phase 2 report provides identification of particular strains, genotypes, serotypes, or lineages as well as detection statistics, including genome coverage histograms (available for any user-selected data point appearing in the plot) to enable a detailed assessment and comparisons of identified viruses (Fig. 2A). For 16 specimens, the viruses were clearly indicated in the phase 1 report. Overall, clear information was obtained for most specimens (Table 1), and strong virus signals (both percent genome coverage and maximum coverage depth) were observed (see Fig. S3 in the supplemental material), providing results ready for interpretation by microbiologists. Furthermore, all of the viruses that were detected by specific RT-PCR in specimen 20 by routine screening (PeV, PIV-1, PIV-3, PIV-4, and MeV) were also highlighted in the phase 1 report (Table 1 and Fig. 2B), demonstrating the capacity of this pipeline to identify coinfections (all detected viruses have genome coverage of ≥83%) and a relatively strong signal (maximum coverage of ≥70%). Specimen 20 (NPA) was collected from a 10-month-old child 8 days postvaccination for measles. In agreement with the traditional Sanger sequencing of the MeV virus N gene, the phase 2 analysis reported the vaccinal Edmonston strain as the most likely candidate (Fig. 2B).

Two specimens (specimens 12 and 17) had multiple strong signals that made it difficult to pinpoint one specific virus in phase 1. However, phase 2 reports and histograms made it possible to distinguish the target virus over other background viruses. For example, specimen 17 was positive for influenza A virus but negative for human rhinovirus by specific RT-PCR (routine laboratory screening), yet HRV represented the strongest signal in the phase 1 report for specimen 17 (Fig. 2C). The phase 2 cross-contamination analysis (CCA) can help to clarify results in such situations. In the CCA bar plot shown in Fig. 2C, neighboring specimen 16 was shown to contain 1,000 orders of magnitude more of the same HRV genotype, HRV-66. As these two specimens were prepared alongside each other in the same run of experiments, the pipeline indicated that specimen 17 was most likely contaminated by specimen 16. While we could not definitively exclude the possibility of coinfection with the same virus genotype, the lower signal observed for specimen 17 made this highly improbable (Fig. 2C). Of note, the robustness of the phase 2 report (presence of the HRV-66 genotype in specimen 16) was confirmed by classical sequencing methods based on an analysis of the VP4-VP2 region and 5′ untranslated region (UTR) (data not shown). The same approach could be used to rule out the very weak signals from JCV in specimen 01, HRV in specimen 10, and measles virus in specimen 12 (see Fig. S4 in the supplemental material). Of note, while these cross-contaminating virus signals were extremely low (only 2 to 8 total mapped reads per virus) and could be considered background signal, the CCA module was still able to detect them.

The ezVIR results from two specimens (specimens 03 and 04) were inconclusive. These cases were CSF and BAL specimens for which VZV and CMV (both herpesviruses) were detected in routine analysis, respectively. While the presence of these herpesviruses was detected in phase 1 reports, the low depth and percent genome coverage limited our ability to determine the exact type of herpesvirus present in the specimen with the current version of our pipeline (see Fig. S3 in the supplemental material). However, the ezVIR phase 1 reports for both of these specimens enabled one to presume that herpesvirus was present—valuable information for physicians nonetheless. Statistically, lower genome coverage can be expected for large (>40,000-bp) virus genomes, such as those of herpesviruses, than for small virus genomes, making the former difficult to highlight. In blood specimens from immunocompromised patients, the interpretation of results can be further hindered by the considerable number of “background” viruses resulting from viral reactivation infections (EBV, TTV, HHV-6, etc.). Despite the relatively large volume (>1,000) of reads mapping to these long genomes, the percent genome coverage and depth of coverage may remain low. This, in turn, can place a detected large-genome virus in the same region as significantly smaller background viruses detected with low depth and percent coverage. In such cases, the larger data point size (which reflects the total genome nucleotides mapped) helps to distinguish the detected target virus from background virus signals (Fig. 1C). Taken together, these observations indicate that improving both the sensitivity and the specificity for *Herpesviridae* is a key point that needs to be addressed in the next version of ezVIR. The “blacklist” option can also help to clarify reports by removing any potential non-clinically relevant viruses that may mask the signal of other detected viruses (for example, removing the TT viruses in...
Specimen-06 positive for Epstein-Barr virus
default report

report with blacklist for TT viruses

<table>
<thead>
<tr>
<th>Genome ID</th>
<th>Genome Name</th>
<th>Short Name</th>
<th>% Genome Cov.</th>
<th>Max Cov. Depth</th>
<th>Mapped reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>gi(428161681)ggi(KC2707813.1)</td>
<td>Human herpesvirus 4 strain Akata_complete genome</td>
<td>Epstein-Barr virus</td>
<td>0.36</td>
<td>24.96</td>
<td>55</td>
</tr>
<tr>
<td>gi(63918161)gi(KX3413.1)</td>
<td>Human herpesvirus 6 strain HHV-6 strain U1102_variant A DNA_complete virion genome</td>
<td>Herpesvirus 6A</td>
<td>0.32</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>gi(463308989)ggi(KC200888.81)</td>
<td>Human herpesvirus 6B_complete genome</td>
<td>Herpesvirus 6B</td>
<td>0.36</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>ENA/FJ751515/FJ751515.1</td>
<td>Micro Torque teno virus_complete genome_isolate microTTV-HD4 2_report 68.125</td>
<td>Micro_TTV_virus</td>
<td>93.62</td>
<td>12342.52</td>
<td>53447</td>
</tr>
<tr>
<td>ENA/J025496/J025496.5</td>
<td>Torque teno virus 19 DNA_complete genome_isolate_TTV_SANBAN_31-MAY-2013</td>
<td>TT_virus</td>
<td>42.25</td>
<td>111050.72</td>
<td>2299068</td>
</tr>
</tbody>
</table>

Specimen-07 positive for CMV
default report

report with blacklist for TT viruses

<table>
<thead>
<tr>
<th>Genome ID</th>
<th>Genome Name</th>
<th>Short Name</th>
<th>% Genome Cov.</th>
<th>Max Cov. Depth</th>
<th>Mapped reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>gi(239909369)ggi(KC1201401.1)</td>
<td>Human herpesvirus 5 transgenic strain Towne_complete genome</td>
<td>Cytomegalovirus</td>
<td>0.43</td>
<td>30</td>
<td>93</td>
</tr>
<tr>
<td>gJ(22347832)ggi(KF793288.1)</td>
<td>Human herpesvirus 1 transgenic strain 171_complete genome</td>
<td>Herpes simplex 1</td>
<td>0.41</td>
<td>18</td>
<td>53</td>
</tr>
<tr>
<td>gi(63918161)gi(KX3413.1)</td>
<td>Human herpesvirus 6 strain HHV-6 strain U1102_variant A DNA_complete virion genome</td>
<td>Herpesvirus 6A</td>
<td>0.51</td>
<td>12.76</td>
<td>45</td>
</tr>
<tr>
<td>gi(463308989)ggi(KC200888.81)</td>
<td>Human herpesvirus 6B_complete genome</td>
<td>Herpesvirus 6B</td>
<td>0.62</td>
<td>11.92</td>
<td>45</td>
</tr>
<tr>
<td>gi(1874225)ggi(NC_001716.2)</td>
<td>Human herpesvirus 7_complete genome</td>
<td>Herpesvirus 7</td>
<td>0.71</td>
<td>13</td>
<td>52</td>
</tr>
<tr>
<td>ENA/FJ751515/FJ751515.1</td>
<td>Micro Torque teno virus_complete genome_isolate microTTV-HD4 2_report 68.125</td>
<td>Micro_TTV_virus</td>
<td>35.53</td>
<td>217.66</td>
<td>250</td>
</tr>
<tr>
<td>ENA/JM12030/JM12030.1</td>
<td>TT_virus strain2007_complete genome__26-MAY-2011</td>
<td>TT_virus</td>
<td>61</td>
<td>10473.56</td>
<td>28150</td>
</tr>
</tbody>
</table>
specimens 06 and 07), as the user can specify viruses that should not appear on the plots (Fig. 3). Of note, regardless of which viruses are displayed on the plots, all analysis metrics for all detected viruses are retained in a corresponding comma-separated data file.

**Dealing with cross-contamination.** Interspecimen contamination is a known consequence of the increased sensitivity of HTS technology (27, 34, 35). Despite exercising the highest degree of precaution during specimen preparation, we nevertheless observed the (albeit weak) presence of viruses (DNA and RNA viruses) from neighboring specimens in 4 specimens (specimen 01 contaminated by specimen 02, specimen 10 by specimen 11, specimen 12 by specimen 13, specimen 17 by specimen 16) (see Fig. S4 in the supplemental material). As previously discussed for specimens 12 and 17, while ezVIR correctly identified the presence of the target viruses, the reports also revealed the presence of viruses not in the original specimens. Cross-contamination during specimen preparation is the most likely cause, as the contaminating virus was always the same as the target virus in neighboring specimens (see Fig. S4). A clear example is the presence of the same HRV-66 genotype present in specimens 16 and 17 (Fig. 2C). It is highly improbable that both patients were independently infected with the same virus, given the large variety of circulating rhinovirus genotypes (11). As these two specimens were prepared alongside each other, the HRV-66 detected in specimen 17 is most likely a result of cross-contamination from specimen 16 rather than the presence of a coinfection. This is supported by the fact that only IAV, not HRV, was identified in specimen 17 by specific RT-PCR as used daily in routine screening. These observations underscore one obstacle that stems directly from the inherent sensitivity of HTS technology that needs to be fully addressed in the future.

**DISCUSSION**

Although various HTS-based virus detection methods currently exist, each is designed to function with a particular HTS platform, and the results remain cryptic for non-experts in bioinformatics. In this proof-of-principle study, a total of 20 clinically relevant positive specimens containing a wide range of viruses (both DNA and RNA) were selected in order to conduct a pilot validation of our HTS-based virus detection tools. The design of this pilot study allowed us to assess the potential effectiveness of using HTS to analyze a representative selection of routine specimens previously characterized by conventional real-time PCR (RT-PCR) (Table 1). While the sample size used in this pilot phase is not sufficient to provide a final validation (i.e., sensitivity and specificity), our results indicate that the success rate and ease of interpretation of results provided by the ezVIR pipeline make it worth considering using HTS as an alternative method for investigation of selected cases. Indeed, 18 of the 20 specimens were correctly analyzed with ezVIR, while 2 specimens remained inconclusive despite the fact that for both specimens, the presence of herpesvirus was clearly detected in phase 1 reports. Nevertheless, as previously mentioned, improving the sensitivity and specificity for *Herpesviridae* members is a key issue that needs to be addressed in the next release of ezVIR. An alternative would be to consider the percentage of similarity and to provide information concerning those reads that specifically correspond to each herpesvirus. Although this pilot study did not validate the sensitivity of HTS data analysis with ezVIR, our results show that it may reach a threshold close to that of real-time PCR. The advantage here is that all known viruses can be detected at once, in contrast to virus-specific PCR methods. The variety of clinical specimens used (BAL fluid, CSF, NPA, plasma, and serum) and the spectrum of viruses tested are representative of those seen in daily practice. In a further step, a larger set of specimens will be needed to corroborate these results. For now, the current pipeline should not be considered a validated diagnostic tool for clinical care.

While the phase 1 report provides rapid species (genus for herpesviruses) identification, the phase 2 report is useful for virus typing (Fig. 2A) and highlighting contamination (e.g., observing an identical strain in multiple specimens with similar patterns in read coverage histograms) and coinfections (Fig. 2B). The reliability of the phase 2 typing reports was demonstrated for 3 specimens that were each confirmed (by classical PCR-based sequencing) to be positive for the correct HRV-66 and MeV B3 genotypes and the MeV Edmonston strain (specimens 16, 13, and 20, respectively). A comparative analysis with a larger set of specimens would be necessary to confirm the robustness of phase 2 typing reports. Although physicians in general may not need typing results for most viruses, this information can be extremely useful in certain situations, such as in the case of immunosuppressed individuals, travelers, vaccinations, or antiviral therapy. An example in this pilot study is the identification of measles virus in a nasopharyngeal aspirate shortly after MeV vaccination with a live vaccine (specimen 20). In such cases, it is important for both physicians and epidemiologists to define whether symptoms are related to the vaccine or circulating strains. The same situation, although rare, is also observed for travelers developing yellow fever meningitis shortly after vaccination.

Due to the intrinsic capacity of HTS to generate millions of sequence reads per specimen, a commonly described consequence of this sensitivity is interspecimen or “environmental” contamination (27, 35). Therefore, current HTS users should be extremely vigilant regarding the viruses present in neighboring specimens analyzed in the same series. Despite implementation of the highest precautions to reduce any potential cross-contamination in this investigation, interspecimen signals were nevertheless observed for four specimens (see Fig. S4 in the supplemental material). However, the presence of such contamination did not affect interpretation of phase 1 reports for two specimens (specimens 01 and 10), and the phase 2 CCA analysis made it possible to correctly interpret results for specimens 12 and 17. When interpreting reports, one must also consider the situation where an unexpected identified virus may indeed be present in the specimen but where the patient was asymptomatic for that particular virus.

The most obvious way to eliminate this issue is to prepare one specimen at a time, a solution not only time consuming but also expensive and impractical for use in routine analysis. Of note, the

**FIG 3** Masking undesired viruses. The “blacklist” option allows users to remove any viruses from plots. This is useful in situations where high levels of irrelevant virus might reduce visibility of other viruses present in the specimen. For example, the human TT viruses found in these clinical cases are not relevant and can be removed with this option. By default, a corresponding table is created with each graphical report to list all viruses found in the specimen, regardless of whether they are blacklisted. Even if a virus is not clearly visible in the default report, it is easily found in the corresponding table. Cov., coverage.
protocols used in HTS technology are based mainly on individual methods and utilize manual kits. The optimization of all such presequencing procedures (sample preparation, nucleic acid extraction, and library preparation) is a key issue that will likely be highly improved by automation. While solutions like this currently exist, they remain cost prohibitive for most laboratories. Such automated solutions may also strongly minimize the contamination detected by HTS methods. In the future, it will be important to define the sample preparation methods best adapted for each routine application (e.g., DNA versus RNA viruses). Currently, the issue of contamination may be partially resolved by considering the virus identification in the context of results from the neighboring specimens, as demonstrated with the ezVIR cross-contamination analysis (see Fig. S4 in the supplemental material). The CCA aims to facilitate the identification of such contaminants based on significant differences in signal strength (percent coverage and depth of coverage) of the same virus between specimens. This leads to a paradigm shift in clinical microbiology: a result can still be validated despite background contamination if the bioinformatics pipeline provides reliable analysis tools. The set of specimens used in this pilot study included one documented case of coinfection (specimen 20). Since ezVIR analysis could efficiently detect all viruses previously detected by routine analysis, our results suggest that this pipeline can reach a sufficient level of sensitivity to identify coinfection cases. The next challenge will be to validate and define a minimal threshold (cutoff values) and statistical means to discriminate with certainty between coinfection and cross-contamination. In this respect, the clinical history will also need to be taken into account.

In summary, a close collaboration among infectious disease specialists, clinical microbiologists, bioinformaticians, and HTS platform technology experts allowed us to harness HTS technology for effective and user-friendly virus detection in clinical specimens. The pipeline was designed to identify viruses using a comprehensive manually curated database containing more than 11,000 complete virus genomes, and most importantly, it automatically generates clear and concise (customizable) representations of the results for non-specialists in HTS. This is an important step toward routine use of HTS for clinical virology.

Outlook. While this pilot study was designed around a comprehensive sampling of clinical specimen types and viruses, there are many options for further advancement and validation, as this collection of specimens is obviously neither complete nor 100% representative. The ezVIR pipeline is designed to be modular and customizable (e.g., HTS data from various sequencing platforms can be used as input, and different custom databases can be built from a list of reference nucleotide sequences in a straightforward manner). While HTS is not yet optimal for use in clinical routine diagnostics, we suggest that it soon can be.

ACKNOWLEDGMENTS
We thank Lia van der Hoek (University of Amsterdam, Netherlands) for advice and useful discussions.

This study was supported by the Swiss National Science Foundation (grant 32003B_146993 to L.K.), the Louis-Jeantet Foundation (E.M.Z., L.K., S.C., and O.P.-S.) and the Faculty of Medicine, Geneva.

Part of the computations were performed at the Vital-IT (http://www.vital-it.ch) Center for high-performance computing of the SIB Swiss Institute of Bioinformatics.

REFERENCES
17. Zaki AM, van Boheemen S, Bestebroer TM, Osterhaus AD, Fouchier RA. 2012. Isolation of a novel coronavirus from a man with pneumonia in