ASM | Microbiology Spectrum

| Genomics and Proteomics | Research Article

A collaborative study evaluating the detection of virus-infected cells by transcriptomic high-throughput sequencing

Noémie Deneyer,[1] Pei-Ju Chin,[2] Guillaume Bayon-Vicente,[3] Pascale Beurdeley,[4] Megan H. Cleveland,[5] Anne-Sophie Colinet,[3] Stéphane Cruveiller,[4] Marc Eloit,[4] Shanaz Gilchrist,[6] Noriko Hashiba,[7] Christophe Lambert,[3] Antonio Lembo,[8] Ka-Wai Leong,[9] Brandye Micheals,[9] Sandrine Moreira,[4] Reiko Nakashima,[10] Simone Olgiati,[8] Nasrin Salehi,[9] Qinyu Sun,[10] Kazuhisa Uchida,[7] Olivier Vandeputte,[3] Yuzhe Yuan,[7] Keisuke Yusa,[7] Valeria Maria Zanda,[8] Arifa S. Khan[2]

AUTHOR AFFILIATIONS See affiliation list on p. 21.

ABSTRACT High-throughput sequencing (HTS) is an agnostic virus detection technology that can replace or supplement the conventional adventitious virus assays used for safety testing of biologics. Transcriptomic HTS is particularly suitable for the detection of replicating viruses. We have evaluated the sensitivity of virus detection by transcriptomic HTS using targeted and non-targeted bioinformatics analysis of infected cells spiked into a background of uninfected cells, mimicking material representing a test sample from product manufacturing. Eight laboratories tested a mix of Raji cells (expressing Epstein-Barr virus [EBV] RNAs without virus production) and EBV-negative Ramos cells at ratios ranging from 0.5 to 0.00005. Seven laboratories performed the analysis using different HTS protocols, while the eighth laboratory quantified EBV RNA recovery in all extracted samples by quantifying EBNA-1 by RT-ddPCR assay. All participants detected EBV transcripts at a 0.001 ratio with the targeted analysis and a 0.01 ratio with the non-targeted analysis (i.e., one infected cell in 100 uninfected cells), with four laboratories achieving detection at the 0.0001 ratio for both analysis. The results of the spiking studies demonstrated the capabilities of transcriptomic HTS for adventitious virus detection in cell lines used for the production of biologics, and highlight that optimization in the workflow can improve HTS virus detection.

IMPORTANCE High-throughput sequencing (HTS) is currently recognized as a powerful technology for broad virus detection. While HTS can be considered an alternative method to replace or supplement the conventional adventitious virus detection assays for testing biologics, its routine implementation has been challenging due to the complexities of the HTS workflow, including the matrix of the test sample, selection of HTS methodology (viromic, genomic, or transcriptomic), and variability in the protocols (from sample preparation to bioinformatic analysis). The study demonstrated the performance and capabilities of the transcriptomic HTS approach for the detection of virus-infected cells using different workflows presented by seven laboratories. The study indicated that regardless of the protocol used, all participants detected viral RNA at the 0.01 ratio; however, certain HTS protocols demonstrated better sensitivity of viral detection.

KEYWORDS high-throughput sequencing, adventitious viruses, viral transcriptomics, raji cell line, ramos line, epstein-barr virus, human herpesvirus 4

Ensuring the absence of adventitious viruses is required for the development of safe biological products, including vaccines, gene and cell therapies, and biotherapeutics (1–4). Adventitious viruses may be unintentionally introduced during the manufacturing process due to three main sources of potential contamination: animal-derived

Editor J. J. Miranda, Barnard College, Columbia University, New York, New York, USA

Address correspondence to Arifa S. Khan, arifa.khan@fda.hhs.gov.

Noémie Deneyer and Pei-Ju Chin contributed equally to this article. Noémie Deneyer is listed as co-first-author as the study coordinator. The order of the other authors is alphabetical.

Nasrin Salehi, Ka-Wai Leong, Reiko Nakashima, Qinyu Sun, and Brandye Micheals are employees of Pfizer Inc. that contributed to this paper. Marc Eloit holds shares of PathoQuest. Shanaz Gilchrist is an employee of Sanofi and holds shares in Sanofi. Noémie Deneyer, Anne-Sophie Colinet, Christophe Lambert, and Olivier Vandeputte are employees of GSK and hold shares in GSK. Valeria Maria Zanda, Antonio Lembo, and Simone Olgiati are employees of EMD Serono RBM S.p.A., an affiliate of Merck KGaA, Darmstadt, Germany. FDA and NIST authors are U.S. government employees. Megan H. Cleveland, Pei-Ju Chin, Arifa S. Khan, Pascale Beurdeley, Stephane Cruveiller, Sandrine Moreira, Guillaume Bayon-Vicente, Noriko Hashiba, Kazuhisa Uchida, Yuzhe Yuan, and Keisuke Yusa declare no conflicts of interest.

See the funding table on p. 22.

Received 19 January 2026
Accepted 10 August 2026
Published 22 September 2026

reagents used during production or for cell culture growth; starting material such as the cell substrate; and operator-related contamination (5–7). To mitigate the risk of viral contamination, Good Manufacturing Practices (GMP) have been implemented by manufacturers, alongside testing using conventional *in vivo* and *in vitro* assays as quality control measures in accordance with regulatory recommendations and risk assessment. Nevertheless, albeit rare, cases of adventitious viruses have been reported in biologics (5, 6, 8, 9), highlighting some limitations of the conventional test methods, which include the inability to detect viruses that fail to replicate in the target species or cells used in the assays, and results that rely on the detection by the assay read-out (such as clinical effects in animals or cytopathic effect, hemagglutination, and hemadsorption in cell culture). Additionally, there may be potential interference from the test article due to the inability to neutralize the production virus (1, 4). The capability of high-throughput sequencing (HTS, also referred to as next-generation sequencing [NGS], or massively parallel sequencing [MPS]) for broad virus detection in biological materials was demonstrated by the unexpected detection of porcine circovirus type 1 (PCV-1) in a commercial vaccine (8) and the identification of a novel Sf-rhabdovirus in the Sf9 insect cell line used for baculovirus-expressed products (9), whereas extensive testing using the conventional *in vivo* and *in vitro* methods had failed to detect these viruses.

Due to its capabilities for broad virus detection, HTS with non-targeted bioinformatics analysis can be considered an alternative method to replace or supplement the conventional adventitious virus detection assays (1, 2, 8–16). Compared to the *in vivo* or *in vitro* viral adventitious agent assays, HTS does not rely on the biological properties of a virus but can detect known and novel viruses based on nucleic acid detection. However, HTS cannot distinguish between infectious and noninfectious viruses, and therefore, a positive signal needs to be followed up to assess the risk regarding biosafety (17). In general, HTS has been demonstrated to detect viruses with a better sensitivity than the currently used *in vivo* adventitious virus detection assays (18–20). In addition, while detection of viruses using PCR assays requires prior knowledge of the viral target sequence for assay design, HTS detection does not depend on the availability of sequence information for broadly detecting known and distantly related novel viruses (9).

HTS can be applied through a genomic, viromic, or transcriptomic approach based on the type of test material (17). The genomics approach targets all viral DNAs and RNAs present in intact cells, while the viromics approach targets both intracellular and released encapsidated viral genomes. On the other hand, the transcriptomics approach targets viral RNA transcripts in infected cells. This approach can detect replicating RNA viruses and viral transcripts of DNA viruses. Furthermore, the transcriptomic HTS approach was used to discriminate replicating ssRNA+ viruses from the inactivated viruses, with a sensitivity equivalent to that of RT-PCR (21). In addition, HTS was demonstrated as a suitable alternative strategy to the animal assays for viral safety testing of cell substrates by analyzing ratios of RNA extracts prepared from infected cells in uninfected cells (19). However, whether detection sensitivity would be comparable when different laboratories employ their own established HTS workflows remained a key question.

To address this critical question, the present study was initiated within the Advanced Virus Detection Technologies Working Group (AVDTWG; sponsored by the Parenteral Drug Association), which focuses efforts on considerations for validation, optimization, standardization, and performance of HTS for viral safety of biological products (11). This spiking study involved seven laboratories, each applying their own HTS protocols for the laboratory work and the bioinformatics pipeline, with the aim of evaluating the sensitivity of transcriptomic HTS for detecting viral RNAs in spiked material mimicking adventitious virus contamination in a cell substrate background. HTS performance was evaluated by analyzing different ratios of virus-infected and uninfected cells using Raji cells, which contain Epstein-Barr virus (EBV; also known as human herpesvirus 4) expressing RNAs, and Ramos cells, a human lymphoid cell line free of EBV transcripts. The overall results demonstrated similar detection across all seven laboratories regardless

of using different RNA extraction methods, sequencing platforms, and bioinformatics pipelines. The study indicates that transcriptome HTS can perform reliably across diverse laboratory environments.

MATERIALS AND METHODS

Cell lines and cell spiked sample preparations

Raji (a Burkitt's lymphoblast-like cell line containing integrated Epstein-Barr virus, also known as human herpesvirus 4) (22) and Ramos (an EBV-negative Burkitt's lymphoblast-like cell line) (23) were obtained from the American Type Culture Collection (ATCC). Because of the limited availability of the number of vials in one batch, different batches of Raji and Ramos were received by the different participants (Table S1). However, all the batches were derived from the same master cell bank for each cell line. Once received, cells were stored in liquid nitrogen or at −80°C prior to total RNA extraction, according to each laboratory's protocol.

All participants followed the same protocol for cell handling. Briefly, cells were thawed and transferred into a 1.5 mL sterile microcentrifuge tube in a biosafety cabinet. The original vial was rinsed with 300 µL of nuclease-free phosphate-buffered saline (PBS) and was added to the same tube, which was then centrifuged at 500 × *g* for 2 min at room temperature. The cell pellet was washed twice with 1 mL of sterile, nuclease-free PBS, pH 7.2, without Mg^{2+} (Quality Biological, Cat. #114-056-101 or Gibco, Cat. #14190) by gentle pipetting and resuspended in 1 mL sterile, nuclease-free PBS, pH 7.2, at room temperature to prepare the spiked samples.

The ratios of Raji and Ramos cells were prepared in-house by each of the participants based on the cell numbers per vial provided in the Certificate of Analysis provided by ATCC and indicated in Table S1. Cell ratios were prepared by serial dilutions using sterile, nuclease-free PBS, pH 7.2, as dilution buffer, and those analyzed by the different participants are indicated in Table 1.

Quantification of EBV RNA by RT-ddPCR assay

Total RNA extracted from the spike and control samples by each laboratory was submitted to the National Institute of Standards and Technology (NIST) lab as a central laboratory for quantifying the EBV transcript copy numbers developed by the Center for Biologics Evaluation and Review (CBER) lab. Total RNA aliquots of 20 µL in DNA LoBind tube (Eppendorf, Cat. #022431005) were shipped in dry ice to NIST for ddPCR analysis. The samples were stored at −80°C immediately upon arrival.

EBV transcript copy number was determined by RT-ddPCR assay targeting the EBNA-1 gene using the One-Step RT-ddPCR Advanced Kit for Probes (Bio-Rad, Cat. #1864022). The PCR reaction was set up as follows in a total volume of 22 µL: 5.5 µL of Supermix (4×), 2.2 µL of reverse transcriptase (20 U/µL), 15 mmol/L DTT, 250 nmol/L primer and probe premix for detecting EBV EBNA-1 (Forward: CTCCTACCTGCA ATATCAGG, Reverse: CCATCACCTCCTTCATCTC, Probe: FAM-CATCTCCGTCATCACCCTCCG; Primer: Probe: 250 nmol/L:250 nmol/L), for human porphobilinogen deaminase (PBGD, MIQE sequence: CATCTGGAGTTCAGGAGTATTCGGGGAAACCTCAACACCCGGCTTCGGAAG

TABLE 1 Cell ratios analyzed by study labs

Raji/Ramos ratios	Lab ID
0.5	1, 4, 5, 6, 7
0.1	1, 2, 3, 4, 5, 6, 7
0.01	1, 2, 3, 4, 5, 6, 7
0.001	1, 2, 3, 4, 5, 6, 7
0.0001	1, 2, 3, 7
0.00005	2
0.00001	1

CTGGACGAGCAGCAGGAGTTCAGTGCCATCATCCTGGCAACAGCTGGCCTGCAGCGCATGGGC TGGCACAAC, Assay ID #dCNS530990783, Bio-Rad), 2 µL of total RNA provided by each laboratory, and 9 µL of nuclease-free water (Ambion, Cat. #AM9937). The reaction mix was loaded into a 96-well plate (Eppendorf, Cat. #951020303) and heat-sealed with a foil lid (Bio-Rad, Cat. #1814040) using PX1 PCR plate sealer (Bio-Rad). The reaction mixture was applied to the droplet generator with oil (Bio-Rad, Cat. #1864110) to form oil-in-water droplets. The plate of mixture was sealed and cycled using ProFlex 96-well thermal cycler (Thermo Fisher Scientific, Cat. #4484075) under the following conditions: 1 h at 50°C, 10 min at 95°C, 40 cycles of 30 s at 95°C and 1 min at 56.5°C, followed by enzyme deactivation for 10 min at 98°C and storage at 4°C before reading. The 96-well PCR plate was read on the QX-200 droplet reader (Bio-Rad) using droplet reader oil (Bio-Rad, Cat. #1863004), and the number of FAM and HEX fluorophores, indicating EBV EBNA-1 and human PBGD-positive droplets, were counted. Three independent assays were performed with triplicate samples for each assay. Wells with droplet events over 10,000 were analyzed, and ≥10 positive droplets were used as a cutoff to determine the sample was positive. The signal was considered positive when at least one positive droplet above the negative cutoff was identified in two out of the three repeated tests performed per sample. Human PBGD was used to normalize the EBNA-1 transcripts per cell. A no template control (NTC) was included in the ddPCR assay.

Total RNA extraction and high-throughput sequencing

Each laboratory used its own protocol for total RNA extraction, library preparation, sequencing, and targeted/non-targeted bioinformatic analysis. Table 2 shows the selection criteria used by each laboratory to identify a positive viral signal following targeted/non-targeted bioinformatic analysis. The details of the HTS workflow for each lab are provided below.

Lab 1 prepared the spiking mixture and negative control using culture medium (DMEM; Gibco, Cat. #11995-065). Total RNA was extracted from 250 µL of sample by RNAeasy Plus Mini Kit (QIAGEN, Cat. #74134), following the manufacturer's miRNA extraction protocol with slight modification. Instead of using 50 µL of nuclease-free water to elute once, samples were eluted with 25 µL, followed by 35 µL of nuclease-free water to ensure a higher yield. For genomic DNA cleanup, 50 µL of eluted sample was mixed with 4 µL of RQ1 DNase (Promega, Cat. #M6101) and 6 µL of reaction buffer (10×) and incubated at 37°C for 30 min. Then, 6 µL of STOP buffer was added to the mixture to stop DNase activity. The final cleanup was performed using RNA Clean & Concentrator-25 (Zymo, Cat. #R1017), following the manufacturer's protocol. The column was eluted twice with 30 µL of nuclease-free water, and approximately 50 µL of cleanup product was harvested. The concentration and integrity of total RNA extractions were determined by Victor X2 fluorometry with Ribogreen method (Perkin Elmer) and TapeStation (Agilent), respectively. ddPCR targeting human genomic content was used to confirm the complete removal of genomic content from total RNA samples: 2 µL of sample was assayed by ddPCR Supermix for Probes (No dUTP; Bio-Rad, Cat. #186-3024) with 11 µL of supermix (2×), 1.1 µL of 20× PBGD primer/probe premix (MIQE sequence: GGGGACAAGATTCTTGATACTGCACTCTCTAAGGTAACAACATCT TCCTCCCCAGTTCTTGTCCCCACTCTTCTTTCCTTCCCTGAAGGGATTCACTCAGGCTCTTTCTGT CCGGCAGATTGG; Assay ID# dHsaCNS546172952, Bio-Rad), and 6.8 µL of nuclease-free water (Cat. #AM9937, Ambion). The thermocycling program included the denaturation at 96°C for 10 min, amplification at 94°C for 30 s followed by 56.5°C for 1 min for 40 cycles, enzyme deactivation at 98°C for 10 min, and storage at 4°C. The thermocycler model, droplet generation, and result reading were described in the RT-ddPCR assay section.

The cDNA synthesis and library preparation used 10 µL of the cleaned-up total RNA preparation. The sequencing libraries were generated by TruSeq Stranded Total RNA with Ribo-Zero Human/Mouse/Rat (Illumina, Cat. #20020596) following the manufacturer's protocol. The quality and quantification of libraries were assayed by Agilent Bioanalyzer 2200 with High Sensitivity DNA Kit (Agilent, Cat. #5067-4626). The sequencing was

TABLE 2 Selection criteria used by each laboratory to identify a positive viral signal

Lab	Targeted analysis	Non-targeted analysis
Lab 1	≥1 read and without other non-EBV-associated taxonomies	1. EBV: ≥1 read and without other non-EBV-associated taxonomies 2. Unexpected virus: ≥1 read and without other nonviral-associated taxonomies such as eukaryotic hosts
Lab 2	≥1 read	1. EBV: ≥1 read 2. Unexpected virus: – ≥10 reads observed for a species and ≥300 nucleotides covered on any accession sequence of that species – Not stacked reads
Lab 3	≥1 read (non-spliced aligned length >60 bp and sequence identity >95%)	EBV and unexpected virus: ≥2 reads are assigned to the same virus taxon at the genus level. Sequence identity >95% and longer than 60 bp
Lab 4	≥1 read	EBV and unexpected virus: 1. Signal intensity, minimum length, and coverage distribution using pre-defined cutoffs on the number of viral reads 2. Manual exclusion of non-viral/irrelevant hits considering names and taxa
Lab 5	≥1 read	EBV and unexpected virus: 1. All reads remaining after the taxonomic challenges (best hit against non-viral species) are considered. 2. Manual expertise to invalidate non-significant reads such as stacked reads
Lab 6	≥1 read with alignment length (in both reads in a pair) >50 bp and sequence identity >95%	1. EBV: ≥1 read with alignment length (in both reads in a pair) >50 bp and sequence identity > 95% 2. Unexpected virus: – ≥20 reads with alignment length (in both reads in a pair) >50 bp and sequence identity >95% – The top virus hit assigned to the accession with the highest number of supporting reads
Lab 7	1. Mean composite Phred score of 20 2. Requiring at least 95% sequence identity over at least 80% of the read	For EBV: 1. At least three unique reads 2. Observed coverage has to be at least 25% of the expected coverage times the mean sequence identity For unexpected signals: 1. Mapped to a complete or nearly complete genome 2. Minimum of 8% coverage of the mapped genome 3. Signal spans at least 20% of the mapped genome 4. Not a redundant signal

performed by Illumina HiSeq X sequencer with paired-end, 2 × 151 bp sequencing mode using Hiseq X Reagent Kits (Illumina, Cat. #FC-501-2501). All spiking samples were run individually without pooling.

For targeted analysis, raw FASTQ reads for virus-spiking stocks generated from HTS platforms with paired-end (2 × 151 bp) mode were processed by CLC Genomics Workbench version 22.0.2 (QIAGEN). Unless mentioned, the modules were applied with the default settings. The raw reads were quality control (QC)-trimmed to remove the sequencing adapter, low-quality reads, and the length less than 75 bps by Trim Reads module. The trimmed reads were mapped to EBV reference genome (NCBI Accession V01555.2) using Map Reads to Reference module. The mapping was performed using high-stringency parameters (length fraction = 0.8, similarity fraction = 0.9) to avoid any false-positive calls, while other mappings remained at their default settings. The mapped tracks were subjected to QC for Read Mapping module to obtain the information of genome coverage and sequencing depth. The reads indicating positive hits were extracted and subjected to NCBI BLASTn alignment against nt. The detection of EBV was called positive if more than 10 aligned reads were presented, and the aligned reads had no non-viral taxa screened by BLASTn against NCBI nr/nt database.

For non-targeted analysis, adventitious virus detection was performed using the in-house non-targeted detection pipeline. Briefly, the raw FASTQ reads were trimmed by BBtools BBDuk v38.69 (24), and the adapter list provided by Illumina was imported to the trimmer. The quality matrix of reads before and after trimming was evaluated using FastQC v0.11.8 (25). The host contents were removed from the QC-trimmed reads by HISAT2 v2.2.1 (26) using GRCh38.p14 genome (GCF_000001405.40) as the index. The host-unmapped reads were extracted using SAMtools v1.12 (27), and the paired-end reads were merged using BBTools v38.69 (24). The merged reads were subjected to CD-HIT-EST v4.8.1 (28) to remove the redundant reads sharing 100% similarity. The de-duplicated reads were subjected to BLASTn v2.10.1+ (29) alignment against Unclustered Reference Virus Database (U-RVDB) v24.1 (30, 31) to identify the viral signals. The reference genomes of candidates were extracted, indexed, and subjected to HISAT2 v2.2.1 (26), along with the QC-trimmed, non-collapsed reads for the realignment process. The counterscreening was performed to verify the fidelity of EBV signals. The QC-trimmed reads resulted in the potential hits from HISAT2 realignment were extracted and subjected to BBtools BBDuk v38.69 (24) to remove the low-complexity reads classified as low complexity by computational filters (e.g., CAGGGG repeats). The processed reads were further subjected to BLASTn v2.10.1+ (32) analysis against nt (database downloaded on 8 July 2022). The candidates were confirmed to be true-positive hits if the reads of such candidate were exclusively related to the designated taxonomy, as revealed by the realignment process, with no crosstalk other than viral taxonomies. Within all possible alignments, the accession with the highest mapped read numbers was reported. The top hit was determined based on the highest EBV-related mapped reads. The final candidate lists were generated as tab-separated values (TSV) and imported into Microsoft Excel 2021 for further review and analysis. The detection of EBV hit was called positive if more than 10 aligned reads were presented, and the aligned reads had no non-viral taxa screened by BLASTn against NCBI nt. For non-EBV-related or unexpected hits, the signal was called positive if equal to or more than one read was aligned to targets associated with viral taxonomy without any crosstalk with non-viral taxonomies.

Lab 2 extracted RNA from each sample (250 µL starting material) using QIAGEN RNeasy Mini Kit (QIAGEN, Cat. #74104), according to the manufacturer's instructions, with an elution volume of 30 µL of 1× TE. Following extraction, a DNase treatment was applied to the sample using Turbo DNase (Invitrogen, Cat. #AM2238), following the manufacturer's instructions. Depletion of ribosomal RNA was then performed using Ribo-Zero Plus rRNA Depletion Kit (Illumina, Cat. #20040892), with an elution volume of 10 µL, which was used for cDNA synthesis according to the manufacturer's protocol for SuperScript Double-Stranded cDNA Synthesis Kit (Invitrogen, Cat. #11917010). Note that Qubit quantification of RNA was assayed after rRNA depletion using Qubit RNA BR Assay Kit (Invitrogen, Cat. #Q10210). The final sample was cleaned up and concentrated with AMPure XP beads (Beckman Coulter, Cat. #A63881) according to the manufacturer's recommendations to a final volume of 20 µL. Qubit quantification of cDNA was assayed using the Qubit dsDNA HS Ready-to-Use Kit (Invitrogen, Cat. #Q33231). Libraries were prepared from 1 ng cDNA of each sample using Nextera XT DNA Library Preparation kit (Illumina, Cat. #15031942) according to the manufacturer's instructions. Libraries were denatured to 4 nM according to Illumina's manufacturer's protocol prior to sequencing on the Illumina NextSeq 500/550Dx sequencer with NextSeq 500/550 High Output Kit v2.5 (300 cycles; Illumina, Cat. #20024908) in a paired-end mode. The read length was set to 2 × 150 bp. The sequencing of each sample was performed in a single run, according to Illumina's instructions.

Data were pre-processed by first demultiplexing the raw sequencing reads using bcl2fastq v2.0.1.17 to extract all reads pertaining to each sample (33). QC-trimmed paired-end sequencing reads were first checked using the FastQC v0.11.5 software (https://www.bioinformatics.babraham.ac.uk/projects/fastqc) (25). Next, trimming of adapter/index sequences and low-quality bases was achieved using Trimmomatic v0.36

(34), and reads shorter than 50 bp after trimming were discarded. After this stage, all remaining reads were considered high-quality reads and ready to be analyzed.

For targeted analysis, reads associated with each sample were processed using an internal data analysis pipeline for targeted adventitious virus detection. High-quality reads were aligned using BWA v0.7.12 (35) to a custom sequence database containing the reference human genome assembly GRCh38.p14 (GenBank: GCF_000001405) and human herpesvirus 4 (GenBank: V01555.2). Samtools v1.9 (27), combined with internal scripts, was used to extract essential statistics such as the number of mapped reads per virus and length coverage. The percentage of identity was derived from the number of variants discovered by BCFtools v1.10.2 (36). Results for the different samples were reported in Excel template files summarizing participant results. As the aim is to evaluate the maximum sensitivity of the method for a known, selected virus, no threshold was applied to evaluate positive samples, i.e., only one EBV-associated read was needed to define a sample as positive for EBV.

For non-targeted analysis, reads associated with each sample were processed using an internal data analysis pipeline for adventitious virus detection, based on considerations for optimization of high-throughput sequencing bioinformatics pipelines for virus detection (37). High-quality reads were aligned using BWA v0.7.12 (35) to a sequence database containing both the reference human genome assembly (GenBank: GRCh38.p14) considered the host in this analysis, and RVDB v24.1 (30) to screen for potential virus reads. During counterscreening (37), reads aligned to RVDB were extracted and then aligned to GenBank v250 sequences using BLAST (32) to assess their origin. In an approach similar to TAMER (38), BLAST results were analyzed in a mixture model to estimate the proportion of reads generated by each candidate genome or sequence. The result of the counterscreening is a taxonomical assignment of the reads, consisting mainly of a table with the GenBank accessions of sequences considered present in the sample tested, their description, their taxonomy, the number of reads mapped to those accessions, and the length coverage. Reads assigned to each accession sequence were extracted and realigned only to the assigned accession sequence. Depth of coverage graphs presenting the number of reads covering each nucleotide of reference sequences are then automatically computed for each accession from alignment files. Virus species were considered as having a high likelihood of presence in the samples if ≥10 reads were observed for a species and ≥300 nucleotides were covered on any accession sequence of that species. For each species potentially present, the sequence with the highest number of reads was selected as the top hit. Results for viruses that were not bacteriophages nor virophages were reported in Excel template files collecting participant results.

Lab 3 prepared total RNA from 250 µL of cell mixtures of Raji and Ramos using the RNeasy Mini Kit (QIAGEN, Cat. #74104) and eluted it in 50 µL of ddH_2O (NEB, Cat. #B1500S). The RNA was treated with RQ1 RNase-Free DNase (Promega, Cat. #M6101) at 37°C for 60 min, purified using RNA Clean & Concentrator-25 Kit (Zymo Research, Cat. #R1017), resuspended in 60 µL of ddH_2O, and stored at −80°C until use. RNA concentrations were measured using Bioanalyzer 2100 (Agilent). Ten nanograms of RNA were used for sequencing library preparation using SMART-Seq v4 Ultra Low Input RNA Kit for Sequencing (Takara Bio, Cat. #634889), including ribosomal RNA removal. Libraries were pooled and analyzed on the NovaSeq 6000 using an S4 cell in a paired-end mode. As an internal control, 0.1 fmol of phiX174 DNA was loaded with the libraries. The read length was set to 2 × 150 bp.

Data underwent quality check using FastQC v0.11.8 (25). Preprocessing included trimming low-quality bases and filtering out simple repeats and host sequence reads. Trimming was performed using fastp (39). Host-origin reads were removed using Bowtie 2 v2.5.1 (40) with global alignment and local alignment against the human genome (GRCh38.p14, release GCF_000001405.39).

For targeted analysis, reads were aligned against the reference sequence, EBV strain B95-8 (V01555, 172,281 bp), using Bowtie 2 v2.5.1 (40). Aligned reads underwent

low-complexity filtering using fqtrim v0.9.7 (41), followed by EBV genome coverage calculation with BEDTools v2.31.0 (42). To enable data comparisons across samples and Illumina platforms, normalization was performed using PPM (mapped reads per million trimmed reads). A "valid" alignment to the virus genome was defined as having >95% sequence similarity and a non-spliced aligned region >60 bp, with no similarity to the host sequence. Valid alignments were considered "true" viral signals.

For non-targeted analysis, the preprocessed reads were analyzed for EBV and unexpected virus detection using the following pipeline: (i) mapping to RVDB; (ii) identification of EBV reads; and (iii) validation of unexpected viral-like reads, their RVDB hits, and classifications. The preprocessed reads were mapped to the Clustered RVDB (C-RVDBv22.0, containing 825,901 records) using Bowtie 2 v2.5.1 (40) with the default parameters. Ambiguous sequence reads were removed to increase the certainty of viral sequence origin. Next, annotation of EBV reads was performed. Reads mapped to C-RVDB were realigned against 206 genomes of human gammaherpesvirus 4 (Lymphocryptovirus humangamma4, NCBI:txid3050299) using Bowtie 2 v2.5.1 (40), followed by low-complexity removal using fqtrim. The EBV reference with the highest number of matched reads was identified as the most likely EBV strain present in the samples. EBV genome coverage (%) was calculated as described in the targeted analysis. Notably, only primary alignments were validated due to the multiple repeat regions in the EBV genome.

Next, C-RVDB reads that were not aligned to EBV were considered potential viral-like reads. These reads were further subjected to low-complexity sequence filter using fqtrim (41) and the Bowtie2 (40), filtering out human mRNA (3,001,485 transcripts, https://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/) and bovine genome, bosTau9 (https://hgdownload.soe.ucsc.edu/goldenPath/bosTau9/bigZips/). Additionally, BLASTn (32) searches against the NCBI/nt nucleotide database were performed using the BLASTn plugin in Geneious R10 (10.0.9 for macOS) to identify and remove reads of non-viral origin.

As a result, the remaining reads, considered as viral-specific reads free from host and non-viral origin sequences, were realigned against C-RVDB v22.0. The C-RVDB v22.0 hits were annotated based on their unique accession numbers in NCBI, providing the taxonomic classification. The top hit with the highest number of matched reads was identified as the most likely viral species present. Reads were realigned against the top hit to calculate the EBV genome coverage and the sequence identity. To clarify "true" virus signals, a virus signal was identified when two alignments were presented with the sequence similarity exceeding 95% and a non-spliced aligned region was longer than 60 bp, excluding any similarity with host sequences.

Lab 4 extracted RNA from each sample (200 µL starting material) using Maxwell RSC Simply RNA Cells Kit (Promega, Cat. #AS1390) on the Maxwell RSC Instrument, according to the manufacturer's instructions, with an elution volume of 50 µL of nuclease-free water. The automatic extraction step included a treatment with DNase I Solution provided in the kit. Total RNA concentration was measured by Qubit 4.0 fluorometer using Qubit HS RNA Assay Kit (ThermoFisher Scientific, Cat. #Q32852). RNA integrity was evaluated with 2100 Bioanalyzer using RNA 6000 Pico Kit (Agilent technologies, Cat. #5067-1513). Library preparation was carried out using SMARTer Stranded Total RNA-Seq Kit v2 – Pico Input Mammalian (Takara, Cat. #634418) and its corresponding index: SMARTer RNA Unique Dual Index Kit (Takara, Cat. #634452). AMPure XP beads (Beckman Coulter, Cat. #A63881) were used for library cleanup according to the manufacturer's instructions to obtain a final volume of 20 µL. These steps were performed using a Microlab STAR robotic workstation (Hamilton). Library quantification was measured using Qubit 3.0 with the Qubit dsDNA BR assay Kit (Thermo Fisher Scientific, Cat. #Q32850), while library size was assessed with 2100 Bioanalyzer using DNA 12000 Kit (Agilent technologies, Cat. #5067-1508). Samples were sequenced on Illumina NovaSeq6000 instrument in paired-end mode 2 × 100 bp according to the manufacturer's instructions. Raji and Ramos cell lines were sequenced together using NovaSeq 6000 SP Reagent

Kit v1.5 (300 cycles) (Illumina, Cat. #20028400), while libraries obtained from mixed Raji/Ramos samples at different cell ratios were sequenced together using the NovaSeq 6000 S1 Reagent Kit v1.5 (300 cycles) (Illumina, Cat. #20028317).

Raw sequencing reads were first demultiplexed using bcl2fastq v2.19.1.403 (33) to extract all reads pertaining to each sample. Trimming of adapter/index sequences and trimming of low-quality bases was achieved using fastp v0.19.7 (39); reads shorter than 75 bp after trimming were discarded. After this stage, all remaining reads were considered high-quality reads and ready to be analyzed.

For targeted analysis, reads associated with each sample were processed using an internal data analysis pipeline for targeted adventitious virus detection. High-quality reads were aligned using BWA-MEM v0.7.17-r1188 (35, 43) to the sequence of Human gammaherpesvirus 4 (GenBank: V01555.2). Metrics of coverage, horizontal coverage, and identity were calculated with a set of public tools: Seqkit (v.2.2.0) (44), bedtools (genomeCoverageBed) (v2.30.0) (42), and infoseq (emboss) (v.6.6.0.0) (45). Internal bash scripts were employed for parsing and data collection. Results for the different samples were reported in Excel template files summarizing participant results. The detection of EBV was called positive if more than one aligned read was present.

For non-targeted analysis, reads associated with each sample were processed using an internal data analysis pipeline for adventitious virus detection. High-quality reads were aligned against *Homo sapiens* GRCh38.p14 (GCF_000001405.40) using Bowtie2 (v2.4.5) (40) to allow for the identification and removal of the reads derived from the human genome background. All the mapping reads were filtered by SAMtools (v1.7) (36). Paired FASTQ were regenerated with BEDtools (bamtofastq) (v2.30.0) (42). All the remaining reads were then aligned against C-RVDBv25 with BWA-MEM (v0.7.17-r1188) (35, 43) tool. Mapping reads were filtered using SAMtools (v1.7) (36). Coverage metrics on the reference viral sequences were extracted using custom scripts and a set of public tools, including SeqKit (v2.2.0) (44), BEDtools (genomeCoverageBed) (v2.30.0) (42), and infoseq (emboss) (v6.6.0.0) (45). Viral sequences were filtered based on signal intensity, minimum length, and coverage distribution using pre-defined cutoffs for the number of viral reads and horizontal coverage. All the sequences surpassing those criteria were further filtered to exclude non-viral/irrelevant sequences based on their names and taxa. For each taxonomic family, the sequence with the highest number of reads was selected as the top hit. Top hits were further filtered considering the ratio of the horizontal coverage obtained to the expected viral genome size. Results for viruses were reported in Excel template files collecting participant results.

Lab 5 extracted RNA from each sample (250 µL starting material) using RNeasy Mini Kit (QIAGEN, Cat. #74104), according to the manufacturer's instructions, with an elution volume of 25 µL of RNase-free water. Following extraction, a DNase treatment was applied to the sample using RQ1 DNase (Promega, Cat. #M6101) following the manufacturer's instructions. The sample was then purified using RNA Clean & Concentrator-25 Kit (Zymo, Cat. #R1017), and purified RNA was eluted in 60 µL of RNase-free water. The purified RNA was assessed for quality and quantity using RNA 6000 Nano Kit (Agilent, Cat. #5067-1511) on the Agilent 2100 Bioanalyzer and the Qubit RNA Quantitation High Sensitivity Assay Kit (Thermo Fisher Scientific, Cat. #Q32852) using the Qubit 3.0 instrument, respectively. Libraries were generated using SMARTer Stranded Total RNA-Seq Kit Pico Input Mammalian (Takara Bio, Cat. #635005), following the manufacturer's protocol, using 10 ng of RNA as input per sample. The fragmentation time was adapted based on the RNA Integrity Number (RIN) and DV200 values for each test sample, as determined during the RNA assessment step using the 2100 Bioanalyzer (Agilent). The final library concentrations were assessed for quality and quantity using High Sensitivity DNA Kit (Agilent, Cat. #5067-4626) on the 2100 Bioanalyzer (Agilent) and the Qubit dsDNA Quantitation High Sensitivity Assay Kit (Thermo Fisher Scientific, Cat. #Q32851) using Qubit 3.0 instrument, respectively. All quality assessments were performed following the manufacturer's instructions for each kit. Libraries were denatured to 4 nmol/L according to Illumina's protocol prior to sequencing. A 10%

PhiX was included in each sequencing run. All samples were sequenced using Illumina NextSeq 500/550 sequencer (NextSeq 500/550) and High Output Sequencing Kit v2.5 (150 cycles; Illumina) in single-end read mode. The read length was set to 150 bp (Illumina, Cat. #20024908). Samples were batched in sets of four for sequencing.

Raw sequencing reads were first demultiplexed using bcl2fastq (33) to extract all reads pertaining to each sample. Reads with low-complexity composition were removed with BBDuk (entropy = 0.9, minlength = 60), as part of BBMap v38.76 (24). An in-house tool, DupliQualfilter, was used to remove low-quality sequences and trim adapters. Then, an optional de-hosting step was performed by mapping cleaned reads with BWA v0.7.12 (43) on a host genome downloaded from public databases. Similarly, filtering of rRNA is an optional step performed on RNA libraries by mapping the remaining de-hosted, cleaned reads with BWA v0.7.12 (43) to a curated rRNA database downloaded from the EBI.

For targeted analysis, the reads associated with each sample were processed using a proprietary data analysis pipeline for targeted adventitious virus detection using Bowtie2 v2.4.2 (40) against a database built with targeted genome species. Mapped reads were then submitted to a taxonomic challenge step, which consists of comparing with the reads against a curated version of the global nucleotidic databank from the Standard Data Class of the European Nucleotide Archive (ENA) from EBI using Magic-BLAST v1.5.0 (46). Reads with only non-viral taxonomic assignation among best hits were discarded. Remaining reads are re-aligned on targeted genome species using Bowtie2 (40) with the same parameters used for the initial mapping. Finally, statistics were produced using SAMtools (27) coverage v1.10 and in-house scripts.

For non-targeted analysis, reads associated with each sample were processed using a proprietary data analysis pipeline for agnostic adventitious virus detection. Cleaned filtered reads were assembled using Megahit v1.1 (47). Non-assembled reads were identified by mapping reads back with BWA v0.7.12 (43) on contigs and keeping reads that did not map (singletons). Contigs and singletons are first compared with NCBI BLAST+ v2.7.1 (32) to a proprietary database (ntvrl) built by filtering and curating viral sequences from the nt databank (downloaded from the Standard Data Class [STD] of ENA from EBI). Viral hits were challenged against the whole nt databanks with NCBI BLAST+ (32), and only reads with no non-viral matches among best hits were conserved. Additionally, a rescue step was performed on contigs with no match on ntvrl by comparing these sequences with NCBI BLAST+ (32) to nrvrl, a proteic version of ntvrl. Positive proteic viral matches underwent the same taxonomic challenge as the previously identified contigs and singletons against the nt database and nr, a proteic version of nt. A viral summary was output with final viral hits which survived all taxonomic challenges classified by viral species. Additional statistics were computed using in-house scripts. The top hit was determined based on the highest number of EBV-related mapped contigs or singletons.

Lab 6 extracted RNA from 250 µL of starting materials using Qiagen RNeasy Plus Kit (QIAGEN, Cat. #74134), according to the manufacturer's instructions, with an elution volume of 30 µL of RNase-free water. Following extraction, samples were treated with Promega RQ1 DNase (Cat. #M6101) followed by post-digestion cleanup using RNA Clean & Concentrator-25 (Zymo, Cat. #R1017), according to the manufacturer's instructions. RNA samples were then quantified using Qubit RNA HS Assay Kit (Invitrogen, Cat. #Q32852). One-hundred nanogram of RNA was subjected to ribosomal RNA depletion and library preparation using the Illumina Stranded Total RNA Prep Ligation with Ribo-Zero Plus Kit (Illumina, Cat. #20040525) according to the manufacturer's instructions. The final samples were cleaned up and concentrated with Ampure XP Beads (Beckman Coulter, Cat. #A63881) to a final volume of 15 µL. The concentration and size of libraries were quantified using the KAPA Library Quant Kit for Illumina (Roche, Cat. #KK4824) and High Sensitivity D1000 ScreenTape (Agilent, Cat. #5067-5584) with reagent (Agilent, Cat. #5067-5585), according to the manufacturer's instructions, respectively. Three or four libraries were equally pooled and spiked-in with 2% PhiX according to

Illumina's instructions. The final library mix was diluted to 730 pmol/L according to Illumina's manufacturer's protocol prior to loading onto the NextSeq1000 sequencer with NextSeq 1000/2000 P2 Reagents (300 cycles) v3 kits (Illumina, Cat. #20046813) in paired-end mode. The read length was set to 2 × 150 bp.

The bcl2fastq v3.10.12 was used to generate FASTQ-formatted files and demultiplex reads corresponding to each sample. FastQC v0.12.1 and MultiQC v1.15 were used to assess the quality of sequencing data. Clumpify, which is a part of BBMap v39.01 (https://sourceforge.net/projects/bbmap), was utilized to remove duplication of reads. BBDuk, as a part of BBMap v39.01, was used to perform trimming of adapter and index sequences, as well as low-quality reads and bases. A read length cutoff of 50 bp and a base quality threshold of Q30 were employed in this process. The in-house bioinformatics pipeline was applied to further process the FASTQ files in the targeted and non-targeted analyses.

For targeted analysis, reads associated with each sample were processed using an internal data analysis pipeline. The STAR v2.7.11a (48) was used to align reads to human genome assembly GRCh38 using the default parameters. The unmapped reads were extracted and aligned to Human herpesvirus 4 (GenBank accession no. V01555.2) with nucleotide BLAST v2.10.0 (32). The resulting alignment data were filtered based on the read length and sequence identity, requiring a minimum length of 50 bp and a minimum identity of 95%. Read pairs that passed the filtering criteria were categorized as positive EBV signals. Additionally, SAMtools v1.17 (27) and custom Python code were used to extract alignment and coverage information.

For non-targeted analysis, following the alignment of the high-quality reads to the human genome using STAR, the unmapped reads were extracted and aligned to RVDB v25.0 using nucleotide BLAST to perform an agnostic search for virus sequences. The resulting alignment data were filtered based on the read pair alignment directions, read length, and sequence identity, requiring a minimum length of 50 bp and a minimum identity of 95%. To address the multi-mapped reads issue, a cleaning process was implemented based on the assumption that each read was primarily derived from a single dominant virus in RVDB. Then, the virus accession with the highest number of supporting reads was assigned as a top hit. To enhance the accuracy of results, the virus hits were filtered out based on a finding from a negative control sample which was utilized to obtain background detection information. This filtering process enabled elimination of false-positive hits from the alignment data. The final list of virus hits was curated by applying a threshold of more than 20 reads as a minimum requirement for inclusion in the list. Custom Python codes were used to process the alignment and coverage information.

Lab 7 extracted RNA from each sample (250 µL starting material) using RNeasy Mini Kit (QIAGEN, Cat. #74104) according to the manufacturer's instructions with an elution volume of approximately 50 µL in RNase-free water. Following extraction, a DNase treatment was applied to the sample using Promega RQ1 DNase (Promega, Cat. #M6101), following the manufacturer's instructions. Post-digestion cleanup was done using RNA Clean & Concentrator-25 Kit (Zymo, Cat. #R1017) with an elution volume of 50 µL in DNase/RNase-free water. Depletion of ribosomal RNA was then performed using the Low Input Ribominus Eukaryote System v2 (Thermo Fisher, Cat. #A15027), with an elution volume of 10 µL that was used for cDNA synthesis according to the manufacturer's protocol for the SuperScript Double-Stranded cDNA Synthesis Kit (Invitrogen, Cat. #11917-010). Libraries were prepared from 1 ng cDNA of each sample using Nextera XT DNA Library Preparation Kit (Illumina, Cat. #FC-121-1024) according to the manufacturer's instructions. Libraries were diluted to 0.45 nmol/L according to Illumina's protocol prior to sequencing on the Illumina NextSeq 2000 in a paired-end mode. The read length was set to 2 × 150 bp. The sequencing of all samples was performed in a single run, according to Illumina's instructions.

For the targeted analysis, raw reads were first filtered to remove low-quality entries (mean composite Phred score was below Q20) and low-complexity reads (mean 64 nt-windowed SDUST complexity was below 2.0) (49), both using an in-house proprietary

tool. Then, for each virus used in the study, the reads were mapped to the corresponding reference genome using NCBI BLASTn with an e-value of 1.0×10^{-5}, a minimum read coverage of 80%, and a minimum identity of 80%. The resulting coordinates for the set of matches against the target genome were then converted to SAM format. The SAM file was processed with mpileup (part of the SAMtools suite) (27). The mpileup output was then converted into frequency tables by tallying nucleotide counts and indels. Subsequently, consensus sequences were generated and aligned using MAFFT (50). All of these operations were automated in a single pipeline.

For the non-targeted analysis, raw reads were first filtered to remove low-quality reads and low-complexity reads using a proprietary tool. Six-frame translations of the remaining reads were then compared using a proprietary k-mer-based algorithm against a curated RefSeq viral proteome database, supplemented with records from C-RVDBv27.0 for viral clades poorly represented in RefSeq. Strong matches were routed directly to signal verification (see below). Reads only tentatively matching viral proteomes were fed to a more investigative sub-pipeline, consisting of a host-prescreen, a viral-screen, and a non-viral counterscreen. In brief, that sub-pipeline started by removing reads that strongly matched the host (here the *Homo sapiens* RefSeq genome GRCh38) using MegaBLAST (32). The remaining sequences were compared against a curated reference genome collection, corresponding to the reference proteome collection mentioned above, using BLASTn (32). Those sequences that matched viruses were then counterscreened against archaea, bacteria, fungi, protists, and selected plant and metazoan genomes, all from RefSeq, still using BLASTn (32). A proprietary algorithm then assigned each sequence to its best-fitting taxonomic bin, leveraging both these BLASTn results and phylogenomic information relating the reference viruses to one another (51). The reads passing this challenge, together with the strongly matching reads branching in from the k-mer-based search, were then screened more exhaustively against NCBI's non-redundant nucleotide database (nt), again using BLASTn (32), in order to reconfirm the matches, identify the best-matching viral strain, and map the reads onto the genome.

All of the operations described above were automated in a single pipeline. Given this set of confirmed matches to viral genomes, a decision tree was then applied. For a hit to be reported, it needed to be a match to a metazoan virus. Its genomic coverage had to be at least 8%, its genomic span (leftmost to rightmost coverage coordinates) had to be at least 25%, and the observed genomic coverage multiplied by its mean sequence identity had to be at least 25% of the expected Lander-Waterman coverage (52), based on the number of unique reads. Signals that passed these criteria are described in this paper. Manual analysis of the results categorized signals into known spiked viruses, known additional viruses, and other viruses. The top hit for each of the five viral spikes was reported based on the best match to that species according to genome coverage and sequence identity.

RESULTS

Overall experimental design

The selection of the EBV-expressing Raji cell line was based on a preliminary comparability study using RT-ddPCR assays performed by the CBER lab, in which Raji cells were selected over Daudi cells, a B-cell lymphoma cell line also known to express EBV, because of the higher EBNA-1 expression in Raji cells (Fig. S1). This allowed testing of a range of spiked cell ratios to determine the virus detection limit by HTS.

Cells were obtained from a common source, and cell ratios were prepared following a common protocol (see Materials and Methods). Three cell ratios were generated for testing by all labs: Raji cells diluted in Ramos cells at ratios of 0.1, 0.01, and 0.001. These ratios were identified based on preliminary investigations by the CBER lab using RT-ddPCR and HTS analyses (data not shown). Additional ratios were included by some of the laboratories (Table 1).

Each laboratory used their own protocols for the different steps involved in HTS (described in Materials and Methods). The extracted RNA samples from all laboratories were quantified for EBV EBNA-1 transcripts at the NIST laboratory. Each organization followed its own bioinformatics pipeline to perform targeted and non-targeted analyses. All targeted analyses consisted of an alignment of reads against the EBV reference sequence (NCBI Accession V01555.2), while non-targeted analyses were performed by alignment of reads against a comprehensive reference viral database (described in Materials and Methods). Importantly, for laboratories that included a host removal step in their analysis pipelines, the human genome GRCh38 or hg19 reference sequences were used, as these do not contain EBV in the main FASTA files. For non-targeted analysis, reads that matched the EBV species and passed the participants' quality criteria were reported for all samples. Non-targeted analyses were also used to identify any non-EBV-related (unexpected) viral sequences present in the samples and to determine the sensitivity of virus detection in a sample that may mimic materials potentially tested during the manufacturing process. The participants' criteria to define a positive signal in both targeted and non-targeted analyses are detailed in the "Materials and Methods" section and summarized in Table 2.

Quantification of EBNA-1 transcript

The RT-ddPCR assay quantifying the EBV EBNA-1 transcripts (EBNA-1 hereafter) was initially developed by the CBER lab to confirm the expression of EBV in Raji cells and the absence of EBV in Ramos cells (Fig. S1). The assay was then used by the NIST lab to quantify EBNA-1 in total RNA extracted by each participant using different RNA extraction protocols. The EBNA-1 expression was normalized per cell using the copy numbers of human porphobilinogen deaminase (PBGD) transcripts. The signal was considered positive when at least one positive droplet was identified in two out of the three repeated tests performed per sample.

As shown in Fig. 1, EBNA-1 was detected in 0.1, 0.01, and 0.001 common ratios from the seven laboratories. EBNA-1 was also detected in 0.0001 ratios from Lab 1, Lab 2, and Lab 7, but not from Lab 3. EBNA-1 was not detected in 0.00005 and 0.00001 ratios tested from Lab 2 and Lab 1, respectively. It was noted that RT-ddPCR signals were observed in Lab 4 and Lab 6 in the negative control samples; however, this could be explained by potential laboratory contamination at the extraction step rather than at the RT-ddPCR step, as RT-ddPCR negative controls, consisting of buffer only, were added to each RT-ddPCR run and were all negative. These results showed that, regardless of the extraction protocol used by the seven laboratories, the recovery of EBV in the different cell ratios tested was similar, indicating that extraction methodology was not a source of variability but rather downstream HTS workflows.

Detection of EBV transcripts by HTS

Each laboratory followed its own HTS sample and library preparation workflow and the bioinformatics pipeline to perform targeted and non-targeted analyses of the reads. An overview of the EBV results for the ddPCR (relative EBV EBNA-1 expression) and for HTS targeted and non-targeted bioinformatics analyses is shown in Table 3 (detailed data is included in Table S2). While the number of EBV reads in the targeted analysis was determined based on aligning to the EBV reference genome (NCBI Accession V01555.2), the number of EBV reads in the non-targeted analysis was based on mapping to closely related EBV species. The difference in the number of EBV-mapped reads reported using targeted and non-targeted analyses could be influenced by various factors, such as QC trimming (Lab 1) or different analysis pipelines (Lab 4).

EBV sequences were not detected from the negative control samples for most participants, except Lab 4 and Lab 7. Lab 4's negative sample, which was composed of Ramos cells only, revealed 100 reads mapping to EBV through targeted analysis (details about the negative controls are described in Materials and Methods). Upon investigation, it was found that 98 reads were misaligned or mapped due to low complexity and

TABLE 3 Analysis of EBV-mapped read numbers in control and spiked samples

Lab ID	Cell ratio	ddPCR	Targeted			Non-targeted	
		Relative EBNA-1 expression	No. of QC-trimmed reads	Reads mapped to EBV[a]	% EBV genome covered	No. of QC-trimmed	Reads mapped to EBV species
Lab 1	Neg	0	40,367,188	0	0.00	422,722,508	0
	0.5	0.5903	833,075,528	3,605,770	81.00	500,179,538	507,671
	0.1	0.1753	770,824,208	1,294,119	73.00	393,365,536	92,747
	0.01	0.0183	811,379,422	72,161	39.00	441,858,918	3,863
	0.001	0.0017	773,330,730	7,462	20.00	451,641,820	446
	0.0001	0.0003	859,913,976	550	2.00	506,822,540	40
	0.00001	0	898,095,652	0	0.00	532,508,348	0
Lab 2	Neg	0	117,480,419	0	0.00	117,646,855	0
	0.1	0.0373	352,841,906	11,992	64.80	352,841,906	12,105
	0.01	0.0067	703,151,965	5,170	56.64	703,151,965	5,196
	0.001	0.0003	721,758,206	343	8.68	721,758,206	353
	0.0001	0.0001	532,208,406	68	2.17	532,208,406	68
	0.00005	0	369,271,958	11	0.17	369,271,958	14
Lab 3	Neg	0	190,986,956	0	0.00	190,986,956	0
	0.1	0.0487	196,017,976	104,596	50.75	196,017,976	104,630
	0.01	0.0027	178,609,380	8,001	14.38	178,609,380	7,997
	0.001	0.0001	191,921,488	437	1.92	191,921,488	439
	0.0001	0	179,705,002	37	0.34	179,705,002	37
Lab 4	Neg	0.0017	957,423,310	100	0.16	835,792,342	0
	0.5	2.4057	1,180,551,898	1,694,053	83.85	1,083,595,936	526,076
	0.1	0.8585	955,740,018	481,563	77.72	906,281,612	65,690
	0.01	0.0365	1,080,174,702	35,471	53.48	1,031,536,298	695
	0.001	0.0060	943,973,404	1,843	18.45	908,004,006	0
Lab 5	Neg	0	57,821,872	0	0.00	57,821,872	0
	0.5	0.3286	50,847,293	40,076	77.07	50,847,293	41,712
	0.1	0.0438	68,769,589	6,427	51.57	68,769,589	6,680
	0.01	0.0033	26,952,859	320	10.35	26,952,859	311
	0.001	0.0007	44,814,561	55	2.54	44,814,561	55
Lab 6	Neg	0.0003	87,683,730	0	0.00	87,683,730	0
	0.5	0.1227	67,587,777	24,148	73.39	67,587,777	26,230
	0.1	0.0173	65,940,461	5,176	47.10	65,940,461	5,572
	0.01	0.0010	74,875,848	845	20.48	74,875,848	893
	0.001	0.0009	71,702,627	84	2.35	71,702,627	88
Lab 7	Neg	0	34,788,211	993	0.57	34,788,211	324
	0.5	0.2798	416,645,736	121,041	73.39	416,645,736	68,958
	0.1	0.0400	617,486,806	46,653	67.57	617,486,806	20,637
	0.01	0.0003	425,093,045	327	3.74	425,093,045	107
	0.001	0.0001	574,776,188	48	0.07	574,776,188	0
	0.0001	0.0001	465,680,524	66	0.01	465,680,524	38

[a]EBV accession number V01555 (source NCBI GenBank).

repetitive regions within the EBV genome and were therefore considered irrelevant according to the participant's criteria. The remaining two reads were hypothesized to be either due to sample cross-contamination (especially as the negative sample was already positive by RT-ddPCR) or to index hopping, as the negative control (Ramos, EBV-negative cells) was multiplexed with a sample composed of Raji (EBV-positive) cells. In both cases, the presence of those two reads that could come from cross-contamination did not impact the results of other cell ratios, as a minimum of 695 EBV reads were found in the lowest positive ratio (0.01 by non-targeted analysis). For Lab 7, interestingly, the negative sample (PCR-grade water) was negative by RT-ddPCR but positive by HTS, with 993 reads

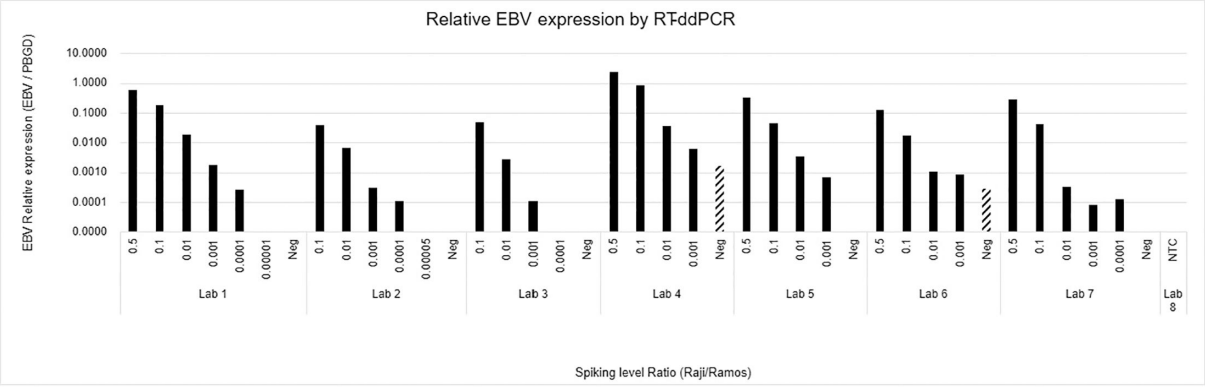

FIG 1 EBV expression analysis of spike-level ratios tested by study labs. The EBNA-1 copy number per µL was determined by ddPCR and normalized with the PBGD copy number per µL to obtain the relative EBV expression. The RT-ddPCR negative samples (Neg), corresponding to buffer only, were included by each participant, and the no-template control (NTC) was included by Lab 8 as the ddPCR assay control for the study. Solid bars indicate positive results in the spiked samples; stripped bars indicate a positive result in the negative control sample.

and 324 EBV reads detected by targeted and non-targeted analyses, respectively. Upon further investigation, it appeared that the presence of EBV in the water control could have resulted from an amplified low-level contamination from a positive sample during sample extraction or library preparation, since reads were clustered in a few locations on the EBV genome. Nevertheless, the analytical results were considered valid by Lab 7 and should not impact the data observed for positive samples.

The results indicated that all seven laboratories successfully detected EBV at cell ratios of 0.1 and 0.01 using both targeted and non-targeted analyses (Table 3), following their respective bioinformatics criteria (Table 2). Moreover, using targeted analysis, the EBV ratio detected by the laboratories was as follows: five at the 0.5 ratio, seven at the 0.001 ratio, four at the 0.0001 ratio, and one at the 0.00005 ratio. EBV was not detected by the single lab that tested the 0.00001 ratio. For non-targeted analysis, similar results were obtained, except that five out of seven detected EBV at the 0.001 ratio. These results demonstrate varying sensitivity of virus detection across different protocols and pipelines used by the participants. A summary table of detection across all spiking ratios is presented in Table 4.

It was noted that all samples that tested positive by RT-ddPCR assay were also identified as positive in the HTS analysis. Interestingly, EBV reads were detected in the 0.00005 and 0.0001 samples, by Labs 2 and 3, respectively, using both targeted and non-targeted analyses, whereas RT-ddPCR results were negative for these samples. This observation might be linked to the detection threshold of the targeted RT-ddPCR assay, which focuses on a segment of the EBNA-1 gene (selected as relevant to EBV latent cycle; [53]), in contrast to sequencing, which can detect genes across the entire EBV genome, including those with higher transcriptional expression than EBNA-1.

The percentage of EBV genome coverage at each cell spiking ratio analyzed by participants is presented in Fig. 2 for the targeted analysis. These results highlight the strong reliability of EBV detection by the targeted analysis. For example, even at the 0.001 ratio, which was the common detection limit of the targeted analysis across seven laboratories, Lab 1 and Lab 4 still achieved approximately 20% coverage.

Across all laboratories and for both targeted and non-targeted methods, a strong linear correlation was observed between the spiked ratio and the relative abundance of EBV, i.e., the ratio between EBV-mapped reads and the total QC-trimmed reads, expressed in reads per million reads (PPM) (Fig. 3). This is also corroborated by the data shown in Table 3. This finding validates the end-to-end workflow, from sample preparation through bioinformatics analysis. The same observation would be made when investigating the ratio between EBV and the host (human) mapped reads, as human reads represent the vast majority of all reads. Furthermore, none of the participants

TABLE 4 Summary of the detection of EBV viral reads in control and test samples using targeted and non-targeted HTS analysis

Test sample[a]	EBV detection[b]	
	Targeted	Non-targeted
Negative control	2/7	1/7
0.5	5/5	5/5
0.1	7/7	7/7
0.01	7/7	7/7
0.001	7/7	5/7
0.0001	4/4	4/4
0.00005	1/1	1/1
0.00001	0/1	0/1

[a]Raji/Ramos cell ratios are indicated for the spiked test samples.
[b]Number of labs that detected EBV/ total number of labs that tested the sample.

exhibited a plateau or ratio resulting in an EBV PPM higher than the previous ratio, suggesting a minimal risk of non-specific EBV detection in the reported results. Note that Lab 7's results were only weakly linear because their method was not optimized for linearity but was instead designed as a limit test.

Detection of unexpected viral sequences in spiked samples

In addition to detecting EBV reads, the non-targeted analysis revealed non-EBV (unexpected) viral sequences, including human endogenous retrovirus sequences, as expected due to the human cell lines used in the spiking study. Additionally, few other viral sequences that were unexpectedly detected may be due to the passage history of the cell lines, kits, and reagents used for sample processing, unknown cross-contamination events, and by misannotation and misclassification of sequences in databases.

Each laboratory reported non-EBV unexpected viral sequences based on the criteria reported in Table 2. Labs 4, 5, and 6 did not report any unexpected viral hits. All unexpected viral hits detected by the non-targeted analysis for each cell ratio analyzed are included in Table S3. For all unexpected signals, the species, the accession number considered as top hit as per participant criteria, and the number of viral reads mapping to the top hit are indicated. For each signal, each participant performed an investigation to justify a true positive signal based on their acceptance criteria. The presence of unexpected viral reads could be explained in some cases due to sequencing platform contamination, known contaminants from commercial kits, or unassigned taxonomy from GenBank sequences. All unexpected viral top hits detected by the non-targeted analysis for the 0.1 ratio are shown in Table 5. It is noteworthy that each laboratory identified unique virus species, with no overlap between labs, indicating a low likelihood of contamination from the Raji and Ramos cell lines.

The results of the non-targeted HTS analysis represented a real-world situation in which broad assays need to be used for adventitious virus testing during the manufacturing of products.

DISCUSSION

The present study used different ratios of EBV RNA-expressing cells spiked into EBV-uninfected cells to demonstrate the potential of HTS transcriptomics for adventitious virus detection in cell lines used for research or in a cell substrate used for manufacturing biologics. The Raji cells with EBV infection were chosen since it is a challenging virus model with only a few genome copies per cell that have been described at the latency phase, with a low level of transcription, whereas the genome copy number and level of transcription of late genes would be higher upon activation to a replicative/lytic stage (54–56).

The impact of different RNA extraction protocols used by the study participants was assessed based on RT-ddPCR analysis on the same test samples that were used for HTS.

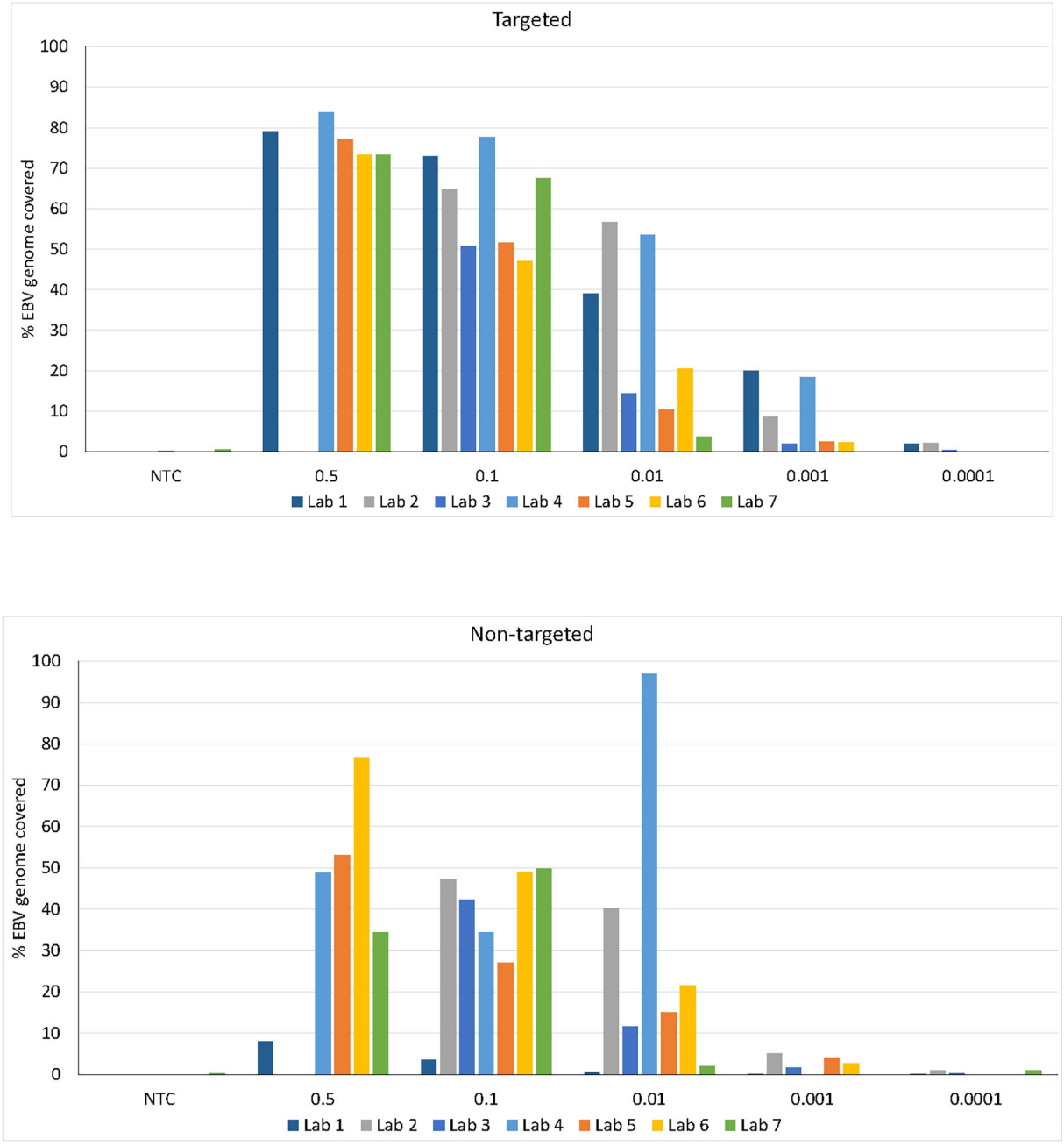

FIG 2 Percentage of the EBV genome covered by HTS targeted and non-targeted analyses. The results are indicated for the cell ratios analyzed by each participant. Labs 2 and 3 did not test the 0.5 cell ratio. Testing of the other cell ratios less than 0.001 is indicated in Table 1.

RT-ddPCR was performed by a central lab to minimize variability for comparing results of the different laboratories. The RT-ddPCR results showed that regardless of the protocols used, the recovery of EBV RNA between the labs was similar for spiked cell ratios ranging from 0.5 to 0.001. However, starting from the 0.0001 ratio and lower, a difference in EBV RNA recovery was observed: although Labs 1, 2, 3, and 7 used the same extraction kit, Lab 3 did not recover EBV RNA (Fig. 1). Similarly, despite the different HTS workflows used by the seven laboratories, all were able to detect EBV transcripts at the 0.1, 0.01, and 0.001 ratios using targeted bioinformatics analysis, or at the 0.1 and 0.01 ratios using non-targeted analysis. Interestingly, EBV was detected by four laboratories that tested the 0.0001 cell ratio and by one laboratory that also tested the 0.00005 ratio. These results demonstrate sequence detection across the virus genome by HTS compared with

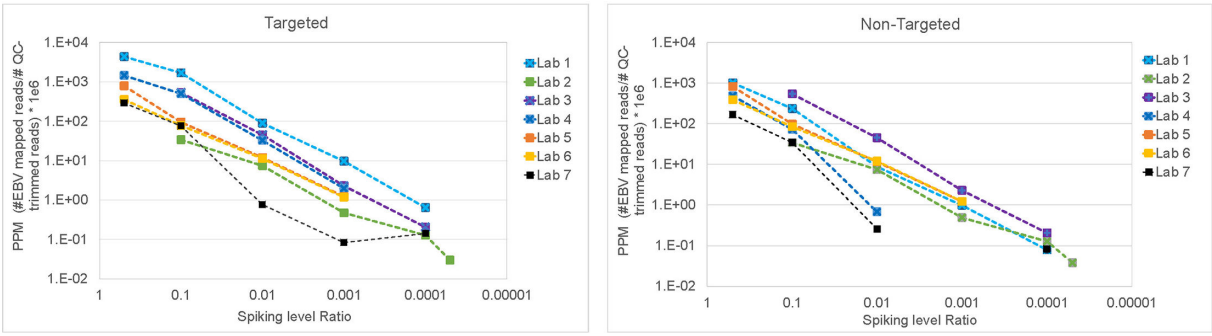

FIG 3 Relative abundance of EBV reads. The results are shown for each spiking level ratio of each participant, for both targeted and non-targeted bioinformatic analyses. Relative abundance is shown as parts per million (PPM) which was computed as the ratio between the number of EBV-mapped reads and the number of QC-trimmed reads.

sequence-specific detection by PCR assay, and further highlight the challenge of virus detection in samples approaching the detection limit.

Differences in the limit of virus detection between the different labs may be due to various factors in the HTS workflow (see details in Table S4): (i) due to the limited cell line availability from the provider, different Raji and Ramos cell batches were used (Table S1). As all the batches were derived from the same master cell bank, no major impact was expected. Results of the ddPCR analysis indicated some differences in EBV expression, but similar cell ratios were detected by labs using different cell lots (0.0001 was detected by Labs 1, 2, and 7; Fig. 1). (ii) Each laboratory followed its own protocol to extract the total RNA from the spiked samples and perform the cDNA synthesis, rRNA depletion, and library preparation. (iii) Sequencing platforms with different throughputs were used, and the multiplexing strategy was different for each laboratory, ranging from

TABLE 5 Unexpected viral hits detected at 0.1 cell ratio by non-targeted bioinformatics analysis[a]

Lab ID	Top viral Hit		
	Accession number of the top hit[b]	Description of the accession number of the top hit	Number of viral reads for the top hit
Lab 1	MH271430.1	Cucumber green mottle mosaic virus isolate CG025, complete genome	2
Lab 1	OM622340.1	*Picornavirales* sp. isolate HPLV-85 genomic sequence	1
Lab 1	MZ824227.1	Reagent-associated tombus-like virus 3 isolate 5 replicase-like gene, partial sequence	8
Lab 1	BR001725.1	TPA_asm: Kadipiro virus rat/2019/399 gene for VP5 protein, complete cds	10
Lab 2	Y18890	Human endogenous retrovirus type K (HERV-K), gag, pol, and env genes	420
Lab 3	KR827417.1	Human picobirnavirus isolate ChXz-6 RNA-dependent RNA polymerase gene, partial cds	4
Lab 3	MF141070.1	Norway luteo-like virus 4 isolate NOR/B1V/Tofte/2014 RNA-dependent RNA-polymerase and coat protein genes, complete cds	29
Lab 3	MT745991.1	Dicistroviridae sp. squirrel/UK/2011 strain UK 2011 genomic sequence, sequence	3
Lab 7	AB047240.1	Human endogenous retrovirus HERV-K(II) DNA, complete sequence, and flanking region	229

[a]Top viral hits are shown. For details, see Table S3, and refer to Table S4 and the "Materials and Methods" section in the paper for the definition of top hit for each participant.
[b]NCBI GenBank Accession numbers.

1 to 21 pooled samples in the same HTS run (yielding a range of reads per sample after trimming and QC of around 45 to 900 million). (iv) Bioinformatics analysis pipelines used for targeted and non-targeted analyses were unique to each lab, including software and parameters used for upstream QC trimming to downstream data analyses. In addition, some pipelines included additional steps, such as host nucleic acid removal, counter screening, and/or taxonomy assignment. Finally, after analysis, each participant applied its own selection criteria (Table 2) to identify the signal reported as viral sequences. For the non-targeted analysis, participants also defined the top hit based on different considerations (e.g., the hit with the highest number of reads, contigs, and singletons, genome coverage, and sequence identity, or the hit with the higher score using NCBI BLAST).

Lab 4 and Lab 7 failed to detect EBV in the 0.001 ratio (the common detection ratio reached for targeted analysis) using non-targeted analysis. However, when employing the targeted pipeline, these laboratories detected 1,843 and 48 EBV reads, respectively. This discrepancy occurred because Lab 4 implemented host DNA removal procedures for non-targeted analysis, but not for targeted analysis. Consequently, reads that mapped ambiguously between the EBV genome and human sequences due to sequence similarity were removed during the host depletion step in non-targeted analysis, while these same reads were retained and detected in the targeted pipeline, where host removal was not performed. The 48 EBV-mapped reads identified by Lab 7 for targeted analysis were rejected in the non-targeted analysis according to Lab 7's criteria. It should be noted that the non-targeted HTS allows for broad virus detection and therefore is best suited for adventitious virus detection in biologics. The results highlighted that non-targeted bioinformatic pipelines still require further optimization to reach sensitivity similar to the targeted approach.

The differences in the selection criteria for a positive result by the non-targeted analysis would impact reporting of the EBV and unexpected (non-EBV) viral hits (Tables S2 and S3, respectively). For example, Lab 2 and Lab 6 reported a positive hit based on at least 10 and 20 mapped reads, respectively, while Lab 1 reported a positive result with one mapped read. In addition, in contrast to the linearity seen with EBV targeted and non-targeted results due to infected Raji cells (Fig. 1), the majority of the unexpected viral signals reported by the non-targeted analysis do not follow the cell ratio, indicating that these sequences may not be due to the Raji and Ramos cell lines (Table S3). Therefore, the unexpected viral signals detected by different laboratories are most likely due to other sources of contamination (e.g., reagents, environmental sources, sequencing platforms, or known contaminants from commercial kits) (57). Except for endogenous retroviral sequences that are normally present in the genome of the cell lines, other unexpected signals were laboratory-specific since they were reported by only one participant. Additionally, virus contamination of cell lines used in the laboratory was reported for a Ramos cell line infected with murine leukemia virus (MLV) (21). This virus was, however, not reported in the Ramos cells used for the present study. Nevertheless, the unexpected viral signals observed during this study highlight the importance of using well-characterized reagents for HTS analysis of a biological sample, especially for method validation.

The detection of background signals due to the laboratory steps and bioinformatics pipelines used in the HTS workflow was revealed by the reporting of EBV signals in the negative control samples by Lab 4 and Lab 7 (Table 3; Table S2). In the targeted analysis of the negative control sample, which comprised of Ramos cells only (described in Materials and Methods), Lab 4 reported 100 reads that were misaligned to EBV. No EBV read was observed in the same negative control sample when applying the non-targeted analysis pipeline. To understand the root cause of the 100 misaligned reads, Lab 4 performed additional investigations. Interestingly, host removal was not integrated into their targeted analysis pipeline. Lab 4, therefore, included a host removal step in its targeted pipeline, leading to a reduction of EBV-mapped reads from 100 to 19 reads. By following the coverage investigation, it has been shown that 17 of the 19

EBV reads aligned to repetitive regions located at nucleotide position 49 kbp of the EBV reference genome, corresponding to the BYRF1 gene in the BamH1 Y/H region (or EBNA-2, NCBI accession number NC_007605), which contains GGGGCA tandem repeats that share similarity with human repetitive elements. The two remaining reads were not removed by the host removal and did not map to a repetitive region (mapped in the *BBRF3* gene). As the negative sample was multiplexed with an EBV-positive sample (EBV expressing Raji cells), the hypothesis for the two remaining reads could be due to index hopping. Index hopping is a phenomenon described by Illumina leading to an incorrect assignment of multiplexed libraries (58, 59). This is an occasional event but can impact sequencing analysis, especially when the depth of sequencing is high. Lab 4 is the participant showing the highest depth of sequencing, with about 1 billion reads after QC and trimming. This high depth of sequencing, coupled with the multiplexing, could have led to the two misaligned EBV reads in the negative sample. Another possibility might be cross-contamination of the negative control at the laboratory level, since a low-level positive signal was detected by RT-ddPCR in the extracted RNA sample from Lab 4.

Lab 7 also reported 993 reads by the targeted analysis that were attributed to EBV in the negative control (PCR-grade water). The sequence coverage for these reads was 0.575% with 99.192% sequence identity. These 993 reads were found to map to six regions of the EBV reference sequence (V01555), with some being read fragments (<150 bp). The first three regions were not detected in the non-targeted analysis because they are non-coding regions, and this lab's current pipeline operates on viral proteomes. Follow-up analysis at the nucleotide sequence level showed that these first three regions did not pass counterscreening in the non-targeted analysis, so they are not viral and are most likely host sequences, as human sequences were detected in the data set. Both the targeted and non-targeted analyses mapped the last three regions to EBV, so this is likely due to cross-contamination from the other five samples that were processed at the same time; a few EBV genomes from these other samples appear to have contaminated the negative control and were amplified during library preparation.

It is important to note that the detection limit reported in this study is specific to EBV expression in the Raji/Ramos cell mix and can therefore not be generalized. However, the approach can be extended to other studies for evaluating viral RNA detection by HTS transcriptomics with the following considerations. For general adventitious virus detection, poly(A)-based enrichment methods should not be used since certain viruses lack poly(A) tails on their mRNAs (e.g., flaviviruses such as Zika, bunyaviruses, and reoviruses). Sensitivity could vary depending on the matrix background, virus transcription levels, and their genome size. EBV presents a large genome (around 176 kb), with multiple genes/transcripts that can increase the detection capacity compared to small virus genomes such as PCV1 (around 1.7 kb). In addition, the choice between splice-aware and non-splice-aware mapping algorithms may impact detection sensitivity. Splice-aware aligners (e.g., STAR, HISAT2) are designed to detect exon-exon junctions by allowing reads to span large gaps corresponding to introns, which is essential for accurate quantification of transcripts with complex splicing patterns. In a pilot comparison performed in one lab using both the splice-aware and non-splice-aware aligners on the same spiking sequencing data sets, there was no significant improvement of detection sensitivity: EBV-mapped read counts and the resulting limit of detection were comparable between the two mapping algorithms, suggesting that, at least for this study, alignment algorithm choice may be less impactful than other factors, such as upstream sample and library preparation strategy, sequencing depth, and downstream bioinformatic analysis. However, this finding may not generalize to viruses with more complex splicing architectures, where splice-aware mapping algorithms may be considered for targeted analyses.

In conclusion, this study used EBV as a model for comparing different HTS transcriptomics workflows to assess the suitability of the approach for detecting viral contaminants in cells and cell-based products. The overall results showed that the HTS

transcriptomics approach is suitable for the detection of transcriptionally active viruses in infected cells in a non-infected cell background. Other approaches would be applied for the detection of transcriptionally inactive or latent viruses. It is important to note that optimization of virus detection and investigation of positive signals still play a major role in the application of HTS for adventitious virus detection, especially using non-targeted pipelines, highlighting the need to design and validate HTS workflows according to the intended application.

ACKNOWLEDGMENTS

PathoQuest thanks Mathilde Mairey, Audrey Boyer, and Louise de Visser for their contributions to the study.

Sanofi acknowledges Robert L. Charlebois for his contributions to the data analysis in the study.

K.U., K.Y., Y.Y., and N.H. were funded by the Japan Agency for Medical Research and Development (AMED) under grant numbers JP24bk0104158 and JP25bk0104193.

A.S.K. conceptualized the study and provided project oversight on the study design, data analyses, and interpretation of the results. N.D. coordinated the study. N.D. and P-J.C. coordinated the data collection, laboratory work, and bioinformatics data analyses from all study participants. G.B-V. performed the laboratory work. C.L. performed bioinformatics data analyses and data interpretation. A-S.C. managed the laboratory work, bioinformatics data analyses, and data interpretation. O.V. provided oversight of laboratory work and data interpretation. V.M.Z. managed the laboratory work, data interpretation, and oversight activities. A.L. performed bioinformatics data analyses and data interpretation. S.O. provided oversight of laboratory work and data interpretation. S.G. performed the laboratory work, bioinformatics data analysis, and data interpretation. B.M. served as the project sponsor and was responsible for securing funding. N.S. provided overall project leadership, including oversight of study design, laboratory execution, data analysis, and interpretation of results. K-W.L. conducted the laboratory experiments and contributed to data interpretation. R.N. and Q.S. performed bioinformatics analyses and participated in the interpretation of the data. N.S. and K-W.L. reviewed the final version for intellectual content and accuracy. M.H.C. performed RT-ddPCR laboratory work and data analysis. M.E. conceptualized the study and provided project oversight on the study design, data analyses, and interpretation of the results. P.B. provided project oversight on the study design and managed the laboratory work. S.M. and S.C. managed bioinformatics data analyses. K.U. managed the project. K.U. and K.Y. supervised the project. Y.Y. performed the laboratory work. Y.Y. and N.H. performed the bioinformatics data analyses.

All authors were involved in drafting and reviewing the manuscript.

AUTHOR AFFILIATIONS

[1]GSK, Wavre, Belgium
[2]Division of Viral Products, Office of Vaccines Research and Review, Center for Biologics Evaluation and Review, U.S. Food and Drug Administration, Silver Spring, Maryland, USA
[3]GSK, Rixensart, Belgium
[4]PathoQuest, Paris, France
[5]Biomolecular Measurement Division, National Institute of Standards and Technology, Gaithersburg, Maryland, USA
[6]Sanofi, Toronto, Ontario, Canada
[7]Graduate School of Science, Technology, and Innovation, Kobe University, Kobe, Japan
[8]EMD Serono RBM S.p.A. Istituto di Ricerche Biomediche A. Marxer, Colleretto Giacosa, Italy
[9]Analytical Research and Development, Microbiology Strategy and Testing Department, Biotherapeutics Pharm. Sci., Pfizer Inc., Andover, Massachusetts, USA

[10]Machine Learning and Computational Sciences, Discovery & Early Development Department, Medicine Design, Pfizer Inc., Cambridge, Massachusetts, USA

PRESENT ADDRESS

Guillaume Bayon-Vicente, Laboratory of Proteomic and Microbiology, UMons, Mons, Belgium
Pascale Beurdeley, Depixus, Paris, France
Stéphane Cruveiller, Censeomics expertise, Yerres, France
Qinyu Sun, Merck & Co., Inc., Rahway, New Jersey, USA

AUTHOR ORCIDs

Noémie Deneyer http://orcid.org/0000-0001-7980-1342
Pei-Ju Chin http://orcid.org/0000-0001-8892-7465
Guillaume Bayon-Vicente http://orcid.org/0000-0002-1486-7647
Christophe Lambert http://orcid.org/0000-0002-7727-2018
Arifa S. Khan http://orcid.org/0000-0002-5633-6131

FUNDING

Funder	Grant(s)	Author(s)
Japan Agency for Medical Research and Development	JP24bk0104158 and JP25bk0104193	Yuzhe Yuan Kazuhisa Uchida Keisuke Yusa Noriko Hashiba

AUTHOR CONTRIBUTIONS

Noémie Deneyer, Project administration, Writing – original draft, Writing – review and editing | Pei-Ju Chin, Data curation, Formal analysis, Investigation, Methodology, Writing – original draft, Writing – review and editing | Guillaume Bayon-Vicente, Methodology | Pascale Beurdeley, Methodology, Supervision | Megan H. Cleveland, Methodology, Writing – original draft | Anne-Sophie Colinet, Supervision, Writing – review and editing | Stéphane Cruveiller, Data curation | Marc Eloit, Formal analysis, Investigation, Supervision, Writing – original draft, Writing – review and editing | Shanaz Gilchrist, Data curation, Formal analysis, Investigation, Methodology, Writing – original draft, Writing – review and editing | Noriko Hashiba, Data curation | Christophe Lambert, Data curation, Formal analysis, Writing – original draft, Writing – review and editing | Antonio Lembo, Data curation, Formal analysis | Ka-Wai Leong, Investigation, Methodology, Writing – review and editing | Brandye Micheals, Funding acquisition | Sandrine Moreira, Data curation, Formal analysis | Reiko Nakashima, Data curation, Formal analysis | Simone Olgiati, Formal analysis, Supervision, Writing – review and editing | Nasrin Salehi, Formal analysis, Investigation, Supervision, Writing – original draft, Writing – review and editing | Qinyu Sun, Data curation, Formal analysis | Kazuhisa Uchida, Funding acquisition, Supervision | Olivier Vandeputte, Formal analysis, Investigation, Methodology, Supervision | Yuzhe Yuan, Data curation, Formal analysis, Investigation, Methodology, Writing – original draft, Writing – review and editing | Keisuke Yusa, Supervision | Valeria Maria Zanda, Investigation, Methodology, Supervision, Writing – original draft, Writing – review and editing | Arifa S. Khan, Conceptualization, Formal analysis, Supervision, Writing – original draft, Writing – review and editing

DATA AVAILABILITY

The HTS raw FASTQ files are available at NCBI BioProject PRJNA1176539.

ADDITIONAL FILES

The following material is available online.

Supplemental Material

Fig. S1 (Spectrum00066-26-S0001.docx). RT-ddPCR results of EBNA-1 expression in Burkitt lymphoma cell lines.
Tables S1 to S4 (Spectrum00066-26-S0002.xlsx). Table S1: Details of cell lots. Table S2: EBV results. Table S3: Unexpected viral hits. Table S4: Comparison of methodologies.

REFERENCES

1. International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). 2023. Viral safety evaluation of biotechnology products derived from cell lines of human or animal origin Q5A(R2).
2. European Pharmacopoeia. 2008. 2.6.16. Tests for extraneous agents in viral vaccines for human use. European Directorate for the Quality of Medicines and HealthCare.
3. European Pharmacopoeia. 2005. Chapter 5.2.3. Cell substrates for the production of vaccines for human use. European Directorate for the Quality of Medicines and HealthCare.
4. US Food and Drug Administration. 2010. Guidance for industry: characterization and qualification of cell substrates and other biological materials used in the production of viral vaccines for infectious disease indications.
5. Barone PW, Wiebe ME, Leung JC, Hussein ITM, Keumurian FJ, Bouressa J, Brussel A, Chen D, Chong M, Dehghani H, et al. 2020. Viral contamination in biologic manufacture and implications for emerging therapies. Nat Biotechnol 38:563–572. https://doi.org/10.1038/s41587-020-0507-2
6. Valiant WG, Cai K, Vallone PM. 2022. A history of adventitious agent contamination and the current methods to detect and remove them from pharmaceutical products. Biologicals 80:6–17. https://doi.org/10.1016/j.biologicals.2022.10.002
7. Petricciani J, Sheets R, Griffiths E, Knezevic I. 2014. Adventitious agents in viral vaccines: lessons learned from 4 case studies. Biologicals 42:223–236. https://doi.org/10.1016/j.biologicals.2014.07.003
8. Victoria JG, Wang C, Jones MS, Jaing C, McLoughlin K, Gardner S, Delwart EL. 2010. Viral nucleic acids in live-attenuated vaccines: detection of minority variants and an adventitious virus. J Virol 84:6033–6040. https://doi.org/10.1128/JVI.02690-09
9. Ma H, Galvin TA, Glasner DR, Shaheduzzaman S, Khan AS. 2014. Identification of a novel rhabdovirus in *Spodoptera frugiperda* cell lines. J Virol 88:6576–6585. https://doi.org/10.1128/JVI.00780-14
10. Khan AS, Ng SHS, Vandeputte O, Aljanahi A, Deyati A, Cassart J-P, Charlebois RL, Taliaferro LP. 2017. A multicenter study to evaluate the performance of high-throughput sequencing for virus detection. mSphere 2:e00307-17. https://doi.org/10.1128/mSphere.00307-17
11. Khan AS, Vacante DA, Cassart J-P, Ng SHS, Lambert C, Charlebois RL, King KE. 2016. Advanced Virus Detection Technologies Interest Group (AVDTIG): efforts on high throughput sequencing (HTS) for virus detection. PDA J Pharm Sci Technol 70:591–595. https://doi.org/10.5731/pdajpst.2016.007161
12. Ng SH, Braxton C, Eloit M, Feng SF, Fragnoud R, Mallet L, Mee ET, Sathiamoorthy S, Vandeputte O, Khan AS. 2018. Current perspectives on high-throughput sequencing (HTS) for adventitious virus detection: upstream sample processing and library preparation. Viruses 10:566. https://doi.org/10.3390/v10100566
13. Mee ET, Preston MD, Minor PD, Schepelmann S, Participants CSS. 2016. Development of a candidate reference material for adventitious virus detection in vaccine and biologicals manufacturing by deep sequencing. Vaccine 34:2035–2043. https://doi.org/10.1016/j.vaccine.2015.12.020
14. Barone PW, Keumurian FJ, Neufeld C, Koenigsberg A, Kiss R, Leung J, Wiebe M, Ait-Belkacem R, Azimpour Tabrizi C, Barbirato C, et al. 2023. Historical evaluation of the *in vivo* adventitious virus test and its potential for replacement with next generation sequencing (NGS). Biologicals 81:101661. https://doi.org/10.1016/j.biologicals.2022.11.003
15. European Pharmacopoeia. 2017. 5.2.14. Substitution of *in vivo* methods by *in vitro* methods for the quality control of vaccines. European Directorate for the Quality of Medicines and HealthCare.
16. Chin P-J, Lambert C, Beurdeley P, Charlebois RL, Colinet A-S, Eloit M, Gilchrist S, Gilleece M, Hess M, Leimbach A, de Man TJB, Vandeputte O, Walas D, Wang W, Khan AS. 2025. Virus detection by short read high throughput sequencing in a high virus low cellular background. NPJ Vaccines 10:61. https://doi.org/10.1038/s41541-025-01104-1
17. European Pharmacopoeia. 2024. 2.6.41. High-throughput sequencing for the detection of viral extraneous agents.
18. Charlebois RL, Sathiamoorthy S, Logvinoff C, Gisonni-Lex L, Mallet L, Ng SHS. 2020. Sensitivity and breadth of detection of high-throughput sequencing for adventitious virus detection. NPJ Vaccines 5:61. https://doi.org/10.1038/s41541-020-0207-4
19. Beurdeley-Fehlbaum P, Pennington M, Hégerlé N, Albert M, Bennett A, Cheval J, Clark A, Cruveiller S, Desbrousses C, Frederick J, Gros E, Hunter K, Jaber T, Gaiser M, Jouffroy O, Lamamy A, Melkowski M, Moro J, Niksa P, Pillai S, Eloit M, Ruppach H. 2023. Evaluation of a viral transcriptome next generation sequencing assay as an alternative to animal assays for viral safety testing of cell substrates. Vaccine 41:5383–5391. https://doi.org/10.1016/j.vaccine.2023.07.019
20. Chin P-J, Tsou J-H, Armstrong A, Deneyer N, Zanda V, Ayama-Canden S, Colinet A-S, Fuentes SM, Korokhov N, Lambert C, Lavorgna A, Noll M, Olgiati S, Protz M, Shaid S, Sohrabi A, Whiteman M, Khan AS. 2025. Evaluation of high-throughput sequencing for replacing the conventional adventitious virus detection assays used for biologics. NPJ Vaccines 11:28. https://doi.org/10.1038/s41541-025-01351-2
21. Brussel A, Brack K, Muth E, Zirwes R, Cheval J, Hebert C, Charpin JM, Marinaci A, Flan B, Ruppach H, Beurdeley P, Eloit M. 2019. Use of a new RNA next generation sequencing approach for the specific detection of virus infection in cells. Biologicals 59:29–36. https://doi.org/10.1016/j.biologicals.2019.03.008
22. Henle G, Henle W. 1966. Immunofluorescence in cells derived from burkitt's lymphoma. J Bacteriol 91:1248–1256. https://doi.org/10.1128/jb.91.3.1248-1256.1966
23. Klein G, Giovanella B, Westman A, Stehlin JS, Mumford D. 1975. An EBV-genome-negative cell line established from an American burkitt lymphoma; receptor characteristics. EBV infectibility and permanent conversion into EBV-positive sublines by *in vitro* infection. Intervirology 5:319–334. https://doi.org/10.1159/000149930
24. Bushnell. 2015. BBMap/BBTools. https://sourceforge.net/projects/bbmap/.
25. Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc.
26. Kim D, Langmead B, Salzberg SL. 2015. HISAT: a fast spliced aligner with low memory requirements. Nat Methods 12:357–360. https://doi.org/10.1038/nmeth.3317
27. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079. https://doi.org/10.1093/bioinformatics/btp352
28. Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658–1659. https://doi.org/10.1093/bioinformatics/btl158
29. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215:403–410. https://doi.org/10.1016/S0022-2836(05)80360-2
30. Chin PJ, Bhavsar JD, Bosma TJ, MacDonald ML, Polson SW, Khan AS. 2025. Refinement of the Reference Viral Database (RVDB) for improving bioinformatics analysis of virus detection by high-throughput

sequencing (HTS). mSphere 10:e0028625. https://doi.org/10.1128/msphere.00286-25
31. Goodacre N, Aljanahi A, Nandakumar S, Mikailov M, Khan AS. 2018. A Reference Viral Database (RVDB) to enhance bioinformatics analysis of high-throughput sequencing for novel virus detection. mSphere 3:e00069-18. https://doi.org/10.1128/mSphereDirect.00069-18
32. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421. https://doi.org/10.1186/1471-2105-10-421
33. Illumina, Inc. 2020. Bcl2fastq conversion software (Version 2.20). https://rrid.site/data/source/nlx_144509-1/search?q=bcl2fastq2&l=bcl2fastq2.
34. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120. https://doi.org/10.1093/bioinformatics/btu170
35. Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25:1754–1760. https://doi.org/10.1093/bioinformatics/btp324
36. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H. 2021. Twelve years of SAMtools and BCFtools. Gigascience 10:giab008. https://doi.org/10.1093/gigascience/giab008
37. Lambert C, Braxton C, Charlebois RL, Deyati A, Duncan P, La Neve F, Malicki HD, Ribrioux S, Rozelle DK, Michaels B, Sun W, Yang Z, Khan AS. 2018. Considerations for optimization of high-throughput sequencing bioinformatics pipelines for virus detection. Viruses 10:528. https://doi.org/10.3390/v10100528
38. Jiang H, An L, Lin SM, Feng G, Qiu Y. 2012. A statistical framework for accurate taxonomic assignment of metagenomic sequencing reads. PLoS One 7:e46450. https://doi.org/10.1371/journal.pone.0046450
39. Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34:i884–i890. https://doi.org/10.1093/bioinformatics/bty560
40. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359. https://doi.org/10.1038/nmeth.1923
41. Pertea G. 2015. Fqtrim: v0.97. https://doi.org/10.5281/zenodo.593893
42. Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842. https://doi.org/10.1093/bioinformatics/btq033
43. Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv. https://doi.org/10.48550/arXiv.1303.3997
44. Shen W, Le S, Li Y, Hu F. 2016. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. PLoS One 11:e0163962. https://doi.org/10.1371/journal.pone.0163962
45. Rice P, Bleasby A, Ison J. 2009. The EMBOSS users guide. http://emboss.open-bio.org/html/use/index.html.
46. Boratyn GM, Thierry-Mieg J, Thierry-Mieg D, Busby B, Madden TL. 2019. Magic-BLAST, an accurate RNA-seq aligner for long and short reads. BMC Bioinformatics 20:405. https://doi.org/10.1186/s12859-019-2996-x
47. Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. 2015. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct *de Bruijn* graph. Bioinformatics 31:1674–1676. https://doi.org/10.1093/bioinformatics/btv033
48. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29:15–21. https://doi.org/10.1093/bioinformatics/bts635
49. Morgulis A, Gertz EM, Schäffer AA, Agarwala R. 2006. A fast and symmetric DUST implementation to mask low-complexity DNA sequences. J Comput Biol 13:1028–1040. https://doi.org/10.1089/cmb.2006.13.1028
50. Kuraku S, Zmasek CM, Nishimura O, Katoh K. 2013. aLeaves facilitates on-demand exploration of metazoan gene family trees on MAFFT sequence alignment server with enhanced interactivity. Nucleic Acids Res 41:W22–w28. https://doi.org/10.1093/nar/gkt389
51. Charlebois RL, Ng SHS, Gisonni-Lex L, Mallet L. 2014. Cataloguing the taxonomic origins of sequences from a heterogeneous sample using phylogenomics: applications in adventitious agent detection. PDA J Pharm Sci Technol 68:602–618. https://doi.org/10.5731/pdajpst.2014.01023
52. Lander ES, Waterman MS. 1988. Genomic mapping by fingerprinting random clones: a mathematical analysis. Genomics 2:231–239. https://doi.org/10.1016/0888-7543(88)90007-9
53. Klein G. 1989. Viral latency and transformation: the strategy of epstein-barr virus. Cell 58:5–8. https://doi.org/10.1016/0092-8674(89)90394-2
54. Arvey A, Tempera I, Tsai K, Chen HS, Tikhmyanova N, Klichinsky M, Leslie C, Lieberman PM. 2012. An atlas of the epstein-barr virus transcriptome and epigenome reveals host-virus regulatory interactions. Cell Host Microbe 12:233–245. https://doi.org/10.1016/j.chom.2012.06.008
55. Wang C, Li D, Zhang L, Jiang S, Liang J, Narita Y, Hou I, Zhong Q, Zheng Z, Xiao H, Gewurz BE, Teng M, Zhao B. 2019. RNA sequencing analyses of gene expression during epstein-barr virus infection of primary b lymphocytes. J Virol 93:e00226-19. https://doi.org/10.1128/JVI.00226-19
56. Majerciak V, Yang W, Zheng J, Zhu J, Zheng ZM. 2019. A genome-wide epstein-barr virus polyadenylation map and its antisense RNA to EBNA. J Virol 93:e01593-18. https://doi.org/10.1128/JVI.01593-18
57. Porter AF, Cobbin J, Li CX, Eden JS, Holmes EC. 2021. Metagenomic identification of viral sequences in laboratory reagents. Viruses 13:2122. https://doi.org/10.3390/v13112122
58. Costello M, Fleharty M, Abreu J, Farjoun Y, Ferriera S, Holmes L, Granger B, Green L, Howd T, Mason T, Vicente G, Dasilva M, Brodeur W, DeSmet T, Dodge S, Lennon NJ, Gabriel S. 2018. Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms. BMC Genomics 19:332. https://doi.org/10.1186/s12864-018-4703-0
59. Illumina. 2017. Effects of index misassignment on multiplexing and downstream analysis. https://www.illumina.com/content/dam/illumina-marketing/documents/products/whitepapers/index-hopping-white-paper-770-2017-004.pdf?linkId=36607862.