

ViroNEXT: A multilayer hybrid system for high-precision detection of known and divergent viruses in metagenomics

Markus Santhosh Braun

Paul-Ehrlich-Institute, Host-Pathogen-Interactions

Gibran Horemheb Rubio Quintanares

Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán

Martin Machyna

Paul-Ehrlich-Institute, Host-Pathogen-Interactions

Janice Durschang

Paul-Ehrlich-Institut

Csaba Miskey

Paul Ehrlich Institut

Xiang-Jun Lu

Columbia University

Pilar Ramos

Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán

Maike Herrmann

Paul-Ehrlich-Institute, Host-Pathogen-Interactions

Pauline Dianne Santos

Paul Ehrlich Institut

Leona Enke

Paul-Ehrlich-Institute, Host-Pathogen-Interactions

Johannes Weydt

Paul Ehrlich Institut

Liam Childs

Paul-Ehrlich-Institute, Host-Pathogen-Interactions

Guillermo M. Ruiz-Palacios

Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán

Walter Ian Lipkin

Columbia University

Johannes Blümel

Paul-Ehrlich-Institut

Renate König

renate.koenig@pei.de

Paul-Ehrlich-Institute, Host-Pathogen-Interactions

software

Keywords: Virus detection, Viral metagenomics, Metagenomic sequencing, Pathogen discovery, Machine learning, Bioinformatics

Posted Date: September 16th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11044166/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

ViroNEXT: A multilayer hybrid system for high-precision detection of known and divergent viruses in metagenomics

Markus Santhosh Braun^{1,7}, Gibran Horemheb Rubio Quintanares^{2,5,7}, Martin Machyna¹, Janice Durschang², Csaba Miskey³, Xiang-Jun Lu⁴, Pilar Ramos⁵, Maike Herrmann¹, Pauline Dianne Santos³, Leona Enke¹, Johannes Weydt⁶, Liam Childs¹, Guillermo M. Ruiz-Palacios⁵, Walter Ian Lipkin⁴, Johannes Blümel^{2,8}, Renate König^{1,8,*}

¹ Paul-Ehrlich-Institute, Host-Pathogen-Interactions, 63225 Langen, Germany

² Paul-Ehrlich-Institute, Department of Infectiology, 63225 Langen, Germany

³ Paul-Ehrlich-Institute, Division Haematology, Cell and Gene Therapy, 63225 Langen, Germany

⁴ Center for Infection and Immunity, Mailman School of Public Health, Columbia University, New York City 10032, NY, USA

⁵ Department of Infectious Diseases, Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán, Mexico City, Mexico

⁶ Paul-Ehrlich-Institute, Information Technology, 63225 Langen, Germany

⁷ co-first author

⁸ co-senior author

*correspondence: renate.koenig@pei.de (R.K.)

Short title: Metagenomic system for viral detection of known and divergent viruses

Keywords

Virus detection, Viral metagenomics, Metagenomic sequencing, Pathogen discovery, Machine learning, Bioinformatics

Abstract

Metagenomic sequencing enables unbiased virus detection but remains challenged by false-positive classifications and limited detection of highly divergent viruses. We present ViroNEXT, a fully automated end-to-end pipeline combining reference-based virus identification with complementary machine-learning-assisted detection of divergent viral sequences. Multilayer nucleotide-, protein-, assembly- and coverage-based analyses improve classification reliability. Across 40 synthetic, spike-in and clinical datasets, ViroNEXT substantially reduced false-positive assignments while achieving recall comparable to or exceeding established workflows and detecting highly divergent viruses missed by homology-based approaches. ViroNEXT provides interpretable automated reporting and is freely available as open-source software and through a public web interface, facilitating broad adoption of viral metagenomic analysis.

Background

Emerging and re-emerging viral pathogens continue to pose a major threat to global public health. Recent outbreaks caused by severe acute respiratory syndrome coronavirus (SARS-CoV-2) and monkeypox virus have highlighted the importance of rapid and unbiased pathogen detection for outbreak response, surveillance and pandemic preparedness^{1,2}. Despite substantial advances in clinical microbiology, the causative agent often remains unidentified in patients with suspected infectious disease³.

PCR- and serology-based assays remain central to routine viral diagnostics. However, these targeted approaches require prior knowledge of the pathogen and may therefore fail to detect highly divergent or previously uncharacterized viruses^{4,5}. Metagenomic next-generation sequencing (mNGS) offers a hypothesis-free alternative by enabling simultaneous analysis of all nucleic acids present within a sample⁴. This approach has already facilitated the identification of numerous viruses in human specimen, e. g. SARS-CoV-2⁶, Merkel cell polyomavirus⁷, astrovirus VA1⁸ and Lujo virus⁹. Beyond clinical diagnostics, agnostic methods for viral detection are also highly relevant for zoonotic surveillance, virus discovery and testing of biological products for viral contamination, where broad-range identification of potential viral contaminants is essential^{10,11}.

Despite its potential, implementation of NGS in routine diagnostics remains technically and computationally challenging. Sequencing datasets generated from clinical or environmental samples frequently contain millions of reads dominated by host-derived nucleic acids, requiring efficient separation of pathogen-derived sequences from host and environmental background signals^{3,11}. Established bioinformatic pipelines typically achieve similar sensitivity for closely related viruses but progressively lose sensitivity as evolutionary distance from available reference genomes increases^{1,5,12}.

At the same time, algorithms may generate substantial numbers of false-positive viral assignments arising from spurious alignments, taxonomic ambiguities, or overly permissive classifications, complicating interpretation of metagenomic findings.

A broad range of metagenomic analysis platforms has been developed in recent years, but many require substantial technical and computational expertise, dedicated infrastructure or commercial subscriptions. Moreover, interpretation of viral classifications frequently requires expert knowledge because complex samples often yield large numbers of candidate viral assignments of uncertain biological relevance. Furthermore, currently available pipelines are often optimized either for classification of known viruses or for exploratory virus discovery, whereas integrated approaches combining both objectives remain comparatively limited¹⁰.

Machine-learning (ML)–based viral classification approaches have recently gained increasing attention for detecting viral sequences lacking closely related references in available databases. In contrast to strict homology-based methods, ML-based models aim to identify general viral sequence patterns and may therefore retain detection capability even for highly divergent or previously uncharacterized viruses⁵. However, existing ML-based approaches are rarely integrated into end-to-end workflows that combine stringent validation, contamination filtering and clinically interpretable reporting.

Here, we present ViroNEXT, a fully automated hybrid system for viral detection from NGS data that provides robust and readily interpretable species-level virus identifications through the integration of reference-based and machine-learning–assisted analyses. To facilitate practical application, ViroNEXT provides automated reporting and can be executed either locally or through a publicly accessible web interface eliminating subscription requirements and reducing technical barriers to metagenomic viral analysis. Our results demonstrate that integrating machine-learning and homology-based approaches within a rigorous classification system improves detection of divergent viral sequences while substantially reducing false-positive classifications, supporting reliable metagenomic viral detection for pathogen discovery, diagnostics and epidemiological surveillance.

Results

ViroNEXT is a fully automated, end-to-end system that integrates two complementary classification branches within a multilayer classification architecture (Figure 1). The reference-based branch provides sensitive detection of viruses represented in current reference databases through sequential nucleotide, amino acid, and consensus-sequence classification. Reads that remain unclassified after reference-based analysis are subsequently processed by a machine-learning branch, which performs *de novo* assembly followed by classification using a deep learning model trained on generalized viral sequence patterns.

A central design principle of ViroNEXT is that assignments are not reported on the basis of a single classification result. Instead, each assignment must pass sequential classification and filtering implemented within a multilayer architecture before being reported. For the reference-based branch, this comprises primary nucleotide classification, secondary amino acid classification, coverage-based assessment, and tertiary consensus-sequence classification. For the machine-learning branch, candidate contigs are evaluated through read-support thresholds, machine-learning-based viral prediction, and protein-level homology screening. This multilayer classification system is designed to maximize recall for divergent viral sequences while maintaining stringent control of false-positive assignments.

To systematically assess the performance of ViroNEXT, we employed complementary synthetic, experimental spike-in and clinical benchmarking datasets, each designed to probe distinct detection scenarios and enable comparison with established viromics pipelines.

Synthetic dataset

We exploited the flexibility of synthetic data generation to systematically modulate sequence divergence, viral abundance and background complexity under controlled conditions.

The datasets were designed as deliberately challenging scenarios that extend beyond conditions typically expected in real-world samples, thereby stress-testing the detection limits and robustness of ViroNEXT. Three scenarios were modeled: (i) viral coinfection, (ii) an environmental virome and (iii) diagnostic escape.

Viral coinfection scenario

To assess the reference-based component of the pipeline and its tolerance to increasing sequence divergence in known viral taxa, we generated synthetic samples with viral substitution rates ranging from 0% to 25% in 5% increments. Sequence divergence was introduced by applying substitutions across the complete reference genomes without preserving annotated genes or constraining mutations to specific genomic regions. The samples were composed of 70% human reads, 20% reads from opportunistic bacteria and fungi commonly associated with skin, and 10% viral reads (Supplementary Figure 1). The viral reads were generated based on the *First WHO International Reference Panel for Adventitious Virus Detection in Biological Products by High-throughput Sequencing Technologies*¹³, which includes seven viruses spanning a broad range of viral properties (Supplementary Table 1).

Reference-based classification within ViroNEXT correctly assigned all taxa at mutation rates of up to 10% (Figure 2a,b). Moreover, ViroNEXT's multilayer architecture integrating primary nucleotide,

secondary amino acid and tertiary consensus-sequence classification with coverage-based filtering eliminated all false-positives across mutation rates from 0% to 25% (Fig. 2a). As expected, recall declined with increasing sequence divergence. At the highest mutation rate (25%), reference-based classification alone no longer reliably detected highly divergent viral sequences (Figure 2a,b). This decline in recall is an inherent limitation of homology-based approaches and was similarly observed for the benchmarked pipelines, Genome Detective¹⁴, One Codex¹⁵ and SeqtoID, former Chan Zuckerberg ID (CZID) platform³ (Figure 2c).

To overcome this limitation, ViroNEXT incorporates a complementary machine-learning branch designed to detect highly divergent viral sequences. The ML branch continued to detect highly divergent viral sequences even at the highest divergence level, thereby compensating for the loss of recall in the reference-based branch (Figure 2c). VirHunter uses convolutional neural networks to learn local nucleotide sequence patterns allowing viral classification without requiring direct sequence homology to a database reference.⁵ Due to the low number of simulated reads (30), *Betacoronavirus 1* could not be recovered by the ML branch. When the number of simulated reads was increased to 1,000, enabling successful contig assembly, the ML branch detected *Betacoronavirus 1* at a sequence divergence of 25%, demonstrating its ability to recognize highly divergent *Betacoronavirus 1* sequences. The ML branch generated only three false-positive classifications across all evaluated mutation datasets (0%, 15% and 20%, data not shown). Post hoc *blastn* analysis showed that two of the false-positive assignments were attributable to primate endogenous retroviral elements embedded in the human genome, whereas the third originated from fungal background sequences (data not shown).

Despite these isolated ML false positives, ViroNEXT achieved the highest overall balance between recall and precision across increasing sequence divergence, outperforming the benchmarked pipelines in terms of overall F1 score at high mutation rates (Figure 2c).

Environmental virome scenario

In addition to the WHO panel, we challenged the pipeline by further increasing the complexity of the dataset. This dataset consisted of a phage background and 27 viral species infecting eukaryotes and was obtained from LaPierre *et al.* (2019)¹⁶. It was chosen to represent scenario of an environmental sample and consisted of 500,000 paired-end (PE) reads with 25% of the reads being eukaryotic viruses with read numbers between 3 and 73,000 per species. Substitution rate varied from 0.7 to 3.3% and indel rate ranged from 0.3 to 1.7% (Supplementary Figure 2).

All evaluated pipelines achieved high recall, with ViroNEXT, Genome Detective and CZID detecting all 27 species, whereas One Codex detected 26 species. On the contrary, marked differences were

observed regarding precision. While Genome Detective is on par with ViroNEXT (no false-positives), CZID made 10 false-positive predictions (precision = 0.73) and One Codex three (precision = 0.90). F_1 scores of the four pipelines were 1.0 for Genome Detective and ViroNEXT, and 0.84 and 0.93 for CZID and One Codex, respectively (see Supplementary Figure 3).

Together, these results demonstrate that ViroNEXT maintained complete viral detection in a complex environmental virome background while avoiding the additional viral assignments observed with CZID and One Codex.

Diagnostic escape scenario

To assess the combined performance of the reference- and ML-based approaches – particularly at sequence similarities at which reference-based classification showed reduced reliability (Figure 2a,b), we generated synthetic viral genomes spanning increasing evolutionary distances from known reference sequences, corresponding to sequence similarities ranging from 100% to 45% (Supplementary Figure 4, Supplementary Table 2). These datasets extend well beyond the effective range of reference-based classification identified in the initial benchmark and model highly divergent viruses that may also evade sequence-dependent targeted assays such as PCR.

As before, reference-based classification accurately detected closely related sequences but exhibited progressively reduced recall with increasing sequence divergence (Figure 3b). However, by integrating ML-based classification, ViroNEXT consistently detected all five viral taxa across the full range of sequence divergence examined, maintaining optimal recall, precision and F_1 score throughout the experiment (Figure 3c). Comparison with previously published pipelines further demonstrated the advantage of integrating reference-based and machine-learning-assisted classification strategies. ViroNEXT maintained a recall of 1.0 across all evaluated divergence levels, whereas recall decreased to 0.4 for One Codex at a branch-length multiplier of 1 $\times$, for Genome Detective at 4 $\times$ and for CZID at 8 $\times$ (Figure 3c).

Together, these results demonstrate that integration of the ML branch extends the effective detection range of ViroNEXT beyond the sequence-divergence limits of reference-based classification and it also enables detection of highly divergent viruses that were missed by the benchmarked pipelines.

Virus spike-in experiments

To complement the synthetic benchmarks, we analyzed datasets from two independent spiking studies using viruses of known identity and concentration. Their defined viral composition enabled quantitative assessment of classification performance, while the obtained sequencing data contained complexity, from sample background, extraction efficiency, library preparation and sequencing

artifacts. These datasets, therefore, allowed us to evaluate ViroNEXT under more realistic experimental conditions.

WHO International Reference Panel

The spike-in panel consisted of the seven viruses listed in Supplementary Table 1, which were added either to blood plasma or to cell culture medium (DMEM)/adenovirus background (10^9 GE/mL) in concentrations from 10^5 GE/mL to 10^1 GE/mL. Across both sample types, ViroNEXT consistently achieved highest precision and high F_1 scores while maintaining recall comparable to the other evaluated pipelines (Figure 4a and 4b). The improved precision reflected the substantially lower number of false-positive classifications produced by ViroNEXT relative to the other pipelines. ViroNEXT was only surpassed in a single condition, where Genome Detective achieved a slightly higher recall at a viral concentration of 10^3 GE/mL in the adenovirus background dataset (Figure 4b). Limit of detections (LODs) were determined for the Adenovirus dataset. Limit of detection (LOD) was determined for the adenovirus background spike-in dataset (Supplementary Table 3). Across the seven spike-in viruses, ViroNEXT achieved detection limits comparable to those of the benchmarked pipelines. Because detection limits do not account for differences in false-positive assignments, this complementary dimension of classification performance is evaluated separately below (Figure 4d).

Multi-virus spike-in dataset

To assess ViroNEXT performance in a complex multi-virus sample, we analyzed a spike-in dataset of 21 different viruses in human plasma background (Supplementary Table 4). For this complex viral panel, all pipelines achieved relatively high recall (Fig. 4c). Genome Detective (0.94), CZID (0.94) and ViroNEXT (0.92) showed the highest recall values, whereas One Codex achieved a slightly lower recall (0.83). More pronounced differences were observed in precision. One Codex (0.97) and ViroNEXT (0.95) achieved the highest precision, followed by Genome Detective (0.87), whereas CZID produced substantially more false-positive classifications, resulting in a precision of 0.19 (Figure 4c). Notably, all pipelines unexpectedly identified *Orthorubulavirus mammalis* (Parainfluenza 5) that was not known to be amongst the spiked-in viruses panel (Supplementary Table 7). Subsequent virus-specific PCR analysis confirmed that *Orthorubulavirus mammalis* was a real virus contaminant from the *Varicellovirus suidalpha1* stock (data not shown).

False-positive assignment across spike-in datasets

To assess pipeline-specific differences in erroneous viral assignments across experimentally generated spike-in datasets, we compared false-positive counts among all evaluated pipelines. These counts

differed significantly among pipelines (Figure 4d, negative-binomial generalized linear mixed-effects model (GLMM) likelihood-ratio test, $\chi^2(3)=193.1$, $P<2.2\times 10^{-16}$). Relative to ViroNEXT, Genome Detective and One Codex produced 7.0-fold (rate ratio [RR]=0.143, 95% CI 0.054–0.378) and 6.4-fold (RR=0.156, 95% CI 0.058–0.417) more false-positive assignments, respectively. CZID produced approximately 288-fold more false-positive assignments than ViroNEXT (RR for ViroNEXT versus CZID=0.00347, 95% CI 0.0014–0.0086; Holm-adjusted $P<0.001$ for all comparisons). Genome Detective and One Codex showed comparable false-positive rates (Holm-adjusted $P=0.71$). The observed differences were independently confirmed by a non-parametric sensitivity analysis using Friedman's test ($\chi^2(3)=50.34$, $P=6.8\times 10^{-11}$, Supplementary Table 3, Supplementary Table 4, Supplementary Table 5). In conclusion, ViroNEXT achieved the lowest number of false-positive assignments across all datasets (Figure 4d).

In summary, these results demonstrate that ViroNEXT maintains detection performance comparable to benchmarked pipelines in experimentally generated spike-in datasets while providing more stringent control of false-positive classifications.

Novel virus scenario

To assess the ability of ViroNEXT to detect viral sequences from families absent from both the reference database and the ML training data, we generated seven modified versions of our custom-filtered Reference Viral Database (RVDB) sequence database¹⁷ by removing all sequences belonging to one viral family represented in the WHO International Reference Panel (NCBI taxon IDs: 39724, 2946187, 3044472, 11632, 11118, 11244 and 10780, corresponding to *Circoviridae*, *Spinareoviridae*, *Orthoherpesviridae*, *Retroviridae*, *Coronaviridae*, *Pneumoviridae* and *Parvoviridae*, respectively). We then analyzed the WHO spike-in dataset at the concentration of 10^5 GE/mL using each modified database to assess the ability of ViroNEXT to detect novel viruses, defined here as viruses whose entire family was systematically excluded from both the reference database and the ML training data. This represents a deliberately stringent scenario of viral emergence, extending beyond the more common situation where the virus classified as a novel belongs to a known viral family with related sequences already represented in reference databases, and tests the ability of ViroNEXT to detect viruses from previously unrepresented viral families.

As expected, removal of the corresponding viral families from the reference database rendered ViroNEXT's reference-based branch unable to detect the respective virus from the WHO International Reference Panel. However, the machine-learning-assisted branch retained detection capability for

most viral families, despite the absence of closely related reference sequences and corresponding training examples (Table 1). Genome coverage based on alignment of assembled contigs to their respective reference genomes exceeded 65% for five of the seven viruses and reached 90% for three viral families, including complete recovery of the *Circoviridae* genome. Lower genome coverage was observed for *Pneumoviridae* while no *Coronaviridae* virus was detected. The latter could be explained by degradation of the *Betacoronavirus* 1 sample over time (Supplementary Figure 5).

Overall, these results demonstrate the ability of the machine-learning branch to recognize viruses lacking known closely related representatives. This supports its potential to detect viruses from previously uncharacterized viral families in real metagenomic samples.

Clinical scenario

Finally, as a real-world proof-of-concept, we assessed ViroNEXT using datasets generated from human specimen to evaluate its performance under clinically relevant conditions. We analyzed a previously published dataset comprising 13 samples mainly from patients with encephalitis and respiratory infections (Supplementary Table 8)¹⁸ and seven samples from a cohort of patients including plasma and urine samples with suspected arboviral infections (Zika, Dengue, or Chikungunya)¹⁹.

Encephalitis and respiratory infections

ViroNEXT successfully identified the viruses associated with disease in clinical samples obtained from patients with encephalitis or respiratory infections (Supplementary Figure 6). Across the evaluated samples, ViroNEXT achieved a recall of 1.0, a precision of 0.96 and an F_1 score of 0.97. The only unsupported assignments not supported by virological diagnostics corresponded to a human endogenous retrovirus, which very plausibly originated from host genomic material rather than erroneous viral classification (data not shown). Compared with the other pipelines evaluated, ViroNEXT achieved the highest performance across recall, precision and F_1 score in this dataset (Supplementary Figure 6).

Arbovirus cohort

In the cohort consisting of patients with suspected arboviral infection (Zik01), among the four serum samples analyzed, blood-associated viruses, including anelloviruses and parvoviruses, were consistently detected across all samples with high concordance among pipelines (data not shown). In

addition, Enterovirus B was identified in sample MXS_0562 by ViroNEXT, One Codex and CZID, but not by Genome Detective (data not shown).

Across the three urine samples, an *Alphapapillomavirus* was detected in sample MXU_0128 by all pipelines. In sample MXU_0218, a *Betapapillomavirus*, a *Betapolyomavirus* and a *Deltapolyomavirus* were identified by all pipelines except CZID. In the remaining urine sample (MXU_0196), Dengue virus was detected by all evaluated pipelines (Figure 5, full dataset in Supplementary Figure 7).

The total number of viral assignments among the evaluated pipelines differed substantially. On average, One Codex reported 13.4 ± 13.0 (mean $\pm$ SD) viral assignments per sample, whereas CZID identified 10.3 ± 3.1 viral assignments (Supplementary Figure 7). In contrast, Genome Detective generated substantially fewer assignments (2.6 ± 2.4), while ViroNEXT produced 2.0 ± 1.0 assignments (Supplementary Figure 7). The lower number of assignments reported by Genome Detective and ViroNEXT was more consistent with the limited number of viral infections expected based on the clinical context.

These results were further supported by cross-pipeline concordance analysis. Among the viral assignments reported by Genome Detective and ViroNEXT, 67% and 86%, respectively, were supported by at least three independent pipelines (Supplementary Figure 7). Conversely, only 17% of Genome Detective assignments and none of the ViroNEXT assignments lacked support from any other pipeline. In contrast, 76% of One Codex assignments and 75% of CZID assignments were unique to a single pipeline. Accordingly, only 14% of assignments reported by One Codex and CZID were supported by three or more pipelines (Supplementary Figure 7).

Overall, these findings demonstrate that ViroNEXT achieves robust viral classification in datasets from human clinical specimen while minimizing assignments not supported by virological diagnosis. In summary ViroNEXT exhibited the highest cross-pipeline concordance among the evaluated pipelines. Given the number of clinical samples evaluated ($n = 20$), these analyses are presented as proof-of-concept evidence of real-world interpretability, rather than a fully powered diagnostic validation.

Discussion

Robust detection of viral sequences in metagenomic datasets remains challenging because viral diversity is incompletely or inaccurately represented in current reference databases and because many classification workflows generate large numbers of false-positive assignments that complicate interpretation^{12,17}. Here, we present ViroNEXT, a freely available end-to-end bioinformatics platform

that can be executed either locally (<https://github.com/brama108/vironext>) or through a freely accessible web interface (<https://vironext.bihealth.org/>). ViroNEXT combines reference-based classification for sensitive detection of viruses represented in current reference databases with machine-learning assisted viral sequence classification to extend detection to highly divergent viruses, while integrating a multilayer classification system for stringent false-positive control. Across both synthetic and experimental datasets, the principal differences between the evaluated workflows were often not in their ability to detect known viruses, which most pipelines achieved with comparable recall, but rather in the frequency of unsupported viral assignments, highlighting false-positive control as a major determinant of metagenomic classification reliability.

This superior precision was consistently evident across experimental and clinical settings. In the WHO spike-in datasets, ViroNEXT produced no false-positive classifications in 82% of samples, whereas all other evaluated workflows generated at least one false-positive assignment in 85% of cases (Figure 4a, b). CZID and One Codex were particularly prone to reporting false positives; in some samples, the number of erroneous classifications exceeded 100, resulting in precision values below 0.1 (Figure 4a,b,c). These observations are consistent with previous benchmarking studies reporting elevated false-positive rates for both platforms^{3,12}. The importance of stringent false-positive control was further illustrated in the clinical datasets. For the encephalitis and respiratory infection cohorts, all workflows achieved similarly high recall, ranging from 0.88 for Genome Detective to 1.0 for ViroNEXT, consistent with previous reports¹². However, while ViroNEXT produced only a single additional viral assignment, Human endogenous retrovirus K, whose biological origin could not be conclusively excluded (precision = 0.96; Supplementary Figure 6). CZID reported dozens of additional viral assignments (precision = 0.19), most of which lacked support from any of the other evaluated workflows. Similar trends were observed in the Zika cohort, where CZID and One Codex generated substantially larger numbers of viral assignments than Genome Detective and ViroNEXT. Most of these additional detections were unique to a single workflow (~75%), whereas fewer than 20% of Genome Detective assignments and none of the ViroNEXT assignments were unique to a single workflow (Figure 5, Supplementary Figure 7). In this cohort, experimental validation of all unexpected viral assignments by PCR was not feasible due to the high number of signals generated by some workflows and the limited availability of sample material (Supplementary table 8). Therefore, the interpretation of certain detections as likely false positives relied on expected sample composition, biological plausibility, and cross-pipeline concordance, following predefined validation rules. Despite this limitation, the consistently higher cross-pipeline concordance observed for ViroNEXT suggests improved classification reliability. Collectively, these findings indicate that elevated numbers of viral detections are unlikely to reflect true biological complexity alone and may instead arise from reduced classification precision.

The high number of false-positive viral assignments observed in several widely used metagenomic workflows has implications that extend beyond clinical diagnostics. In epidemiological surveillance, spurious viral assignments may distort estimates of virus circulation or prompt unnecessary follow-up investigations, whereas in biosafety testing of biological products they may trigger costly confirmatory testing and complicate the assessment of potential viral contaminants. In pathogen discovery, unsupported classifications may divert resources toward the characterization of apparent viral signals that ultimately cannot be confirmed. By requiring concordant evidence across multiple sequential classification steps, ViroNEXT provides a stringent system for reducing such spurious assignments. Importantly, minimizing false positives does not preclude the detection of genuine, unexpected contaminants. This was exemplified by the unexpected detection of *Orthorubulavirus mammalis* as a true contaminant in the *Varicellovirus suidalpha1* stock, which was subsequently confirmed by PCR. Minimizing false-positive assignments reduces the burden of confirmatory testing and enables focused validation of unexpected viral signals, facilitating the distinction between spurious and genuine findings.

Beyond precision, extending the detection range to highly divergent viruses remains a critical limitation of solely homology-based methods. Consistent with previous studies, the recall of all evaluated workflows, including the reference-based branch of ViroNEXT, declined progressively with increasing sequence divergence³. However, the integration of the ML branch substantially extended this range, enabling the recovery of viral sequences with nucleotide identities as low as 45%—the lower boundary of our benchmarking design. While this represents the lower boundary of our benchmarking design rather than a proven absolute limit of the classifier, it demonstrates the capacity of ViroNEXT to identify highly divergent viral sequences. At these divergence levels, none of the other evaluated workflows retained reliable detection capabilities. Furthermore, leave-one-out analyses demonstrated that the ML branch could detect divergent viruses even when their entire families were excluded from the training data. This extended range likely reflects the capacity of VirHunter’s convolutional neural networks to learn localized nucleotide patterns that distinguish viral from non-viral sequences independent of direct homology, confirming the potential of ML to capture generalized viral characteristics missed by conventional alignment tools^{5,20}.

Nevertheless, ML-assisted detection remains dependent on the successful *de novo* assembly of viral contigs, which in turn requires sufficient sequencing depth and sample quality. This limitation was illustrated by the inability to recover *Betacoronavirus 1* in both the synthetic and experimental WHO datasets. In the synthetic dataset, *Betacoronavirus 1* was represented by only 30 reads within a 30.7 kb genome, substantially limiting *de novo* assembly and downstream contig-based classification. However, when sufficient reads were simulated to enable assembly, the ML branch successfully

detected *Betacoronavirus 1* with 25% sequence divergence (Supplementary Figure 8), confirming that the model is capable of recognizing divergent coronavirus sequences when adequate data are available. In the experimental plasma-background spike-in dataset, *Betacoronavirus 1* exhibited a higher limit of detection (LOD of 10^5 GE/mL) compared to the other viruses in the panel, which were consistently detected at lower concentrations (Figure 4a). This elevated LOD does not reflect an intrinsic limitation of the system for detecting coronaviruses, but rather partial degradation of the viral stock material. This degradation was confirmed by qPCR, which similarly failed to detect *Betacoronavirus 1* below 10^5 GE/mL in that specific sample (Supplementary Figure 5). Together, these observations indicate that the failure to recover *Betacoronavirus 1* resulted primarily from insufficient recoverable viral sequence rather than an intrinsic limitation of the ML classifier.

To ensure these advanced analytical capabilities are accessible for routine application, ViroNEXT was designed to overcome the practical barriers of metagenomics, such as the need for dedicated computational infrastructure and specialized bioinformatics expertise. By integrating automated multilayer classification and standardized, interpretable reporting into a single workflow executable via a freely accessible web interface, ViroNEXT enables researchers and clinical laboratories to perform comprehensive viral screening without extensive computational training.

Conclusions

In conclusion, ViroNEXT integrates complementary reference-based and ML-assisted classification within a multilayer system to extend viral detection beyond the capabilities of conventional homology-based approaches, while maintaining stringent control over false-positive and unsupported assignments. Across synthetic, spike-in, and clinical datasets, ViroNEXT substantially reduces false-positive assignments, extends detection to highly divergent viruses, and achieves recall comparable to or exceeding that of established metagenomic workflows. Although larger clinical studies will be required to establish diagnostic performance, our analyses demonstrate the applicability of ViroNEXT to real-world clinical metagenomic data. By addressing a major barrier to the interpretation of viral metagenomic results, the occurrence of unsupported viral assignments, ViroNEXT provides a robust, practical, and freely accessible system for pathogen discovery, epidemiological surveillance, biosafety testing, and clinical research.

Methods

Pipeline infrastructure & composition

All analyses in this manuscript were performed on the Paul-Ehrlich Institute high-performance computing (HPC) cluster consisting of four fat nodes (each equipped with 64 cores and 1 TB RAM) and two thin nodes (each equipped with 36 cores and 500 GB RAM) using the SLURM workload manager.

The entire ViroNEXT workflow has been implemented as a *snakemake* pipeline to enhance reproducibility, portability and scalability²¹. ViroNEXT is prepared to analyze NGS reads generated by the Illumina short read technology and accepts single-end and paired-end reads alike. The pipeline works with any type of sample origin (e.g. plasma, swabs or environmental samples) and library enrichment strategies (untargeted shotgun-metagenomics and hybridization/capture assays). Jobs can be submitted to a local machine, a high-performance computing (HPC) cluster or executed in the cloud via a web interface available via <https://github.com/brama108/vironext>. Interpretation of pipeline outputs is simplified through interactive HTML reports providing easily understandable results.

ViroNEXT comprises two complementary components: a reference-based and an ML-based classification. Both employ multilayered classification strategies designed to reduce false-positive detections. Rather than relying on a single classification result, candidate viral assignments must pass a stringent sequential classification criteria before being selected (Figure 1A).

Reference-based classification component

Demultiplexed paired-end reads are validated to ensure proper file format using *biopet-validatefastq 0.1.1*²². They undergo preprocessing with *fastq-screen 0.15.3* to remove host sequences²³. *fastq-screen* is used with the `--nohits` flag to ensure the retention of unmapped reads containing potential viral signals.

In case of paired-end reads, the disordered, unmapped data are re-paired for downstream analysis using *repair.sh* from the *BBTools 39.01*²⁴. Quality and adapter trimming are performed in paired-end mode with a quality cutoff of 30 using *cutadapt 5.1*²⁵. The quality of the sequencing data is assessed pre- and post-trimming using *FastQC 0.12.1*²⁶.

Trimmed reads are then mapped against a filtered C-RVDB version 29 using *bwa-mem2 2.2.1*²⁷ (primary classification). While primary classification lays the foundation for maximum recall, the multilayer architecture comprising multiple orthogonal classification strategies successively increases precision.

Multilayer architecture

Alignment files resulting from primary classification are compressed, sorted and indexed using *samtools 1.16.1*²⁸. Quality and reliability of the alignments are ensured using stringent filtering criteria implemented in *sambamba 1.0*²⁹ to eliminate alignments of inferior quality and false-positive classifications, thus enhancing the fidelity of subsequent analysis. Viral abundances are estimated using the compute-abundances script from the *MiCoP* toolkit¹⁶.

Reads aligned in primary classification are re-classified using *Kaiju 1.10.1*³⁰ and RVDB-prot database version 29¹⁷ (secondary classification). To further reduce false-positive assignments, only taxa identified by both the primary and secondary classification steps are retained. Moreover, false-positives are further diminished using stringent coverage filtering implemented in *Python3*. Filters included minimum number of non-adjacent regions (default = 3), a user-specified minimum region length (default = 150 nt) and minimum number of total covered bases (default = 500 nt). This filter excludes assignments supported only by short, localized regions of sequence similarity and is based on de Vries *et al.* (2021)^{31,31}. Trimmed reads are then realigned against taxa passing coverage filtering as mentioned before. Consensus sequences are created using *samtools consensus*²⁸ and classified using *blastn* against C-RVDB version 29¹⁷ (*blast+* 2.7, tertiary classification³²). The best BLAST hit of each consensus sequence was determined using *Python3* based on bitscore, *E*-value and maximum identity. In order to pass tertiary classification, species-level assignments of *blastn* must match the species the identified during primary classification.

The final checkpoint, reported as “high quality read assignment” is an indicator of the proportion of high confident read mappings (mapping quality ≥ 60), providing additional evidence for taxonomic assignments of closely related taxa.

ML-based classification component

This complementary strategy extends detection beyond strict homology-based approaches and increases recall for highly divergent or previously uncharacterized viruses. During reference-based classification, reads that cannot be confidently assigned to any sequence in the reference databases are labeled as unclassified. Such reads may originate from previously undescribed viruses or from highly divergent lineages without representative reference sequences in viral databases. To recover these potentially uncharacterized viral signals, unclassified reads are subjected to *de novo* assembly using *SPAdes 3.15.5* in *rnal* mode³³. The resulting contigs are subsequently classified by *VirHunter 1.0.1*⁵ based on sequence patterns learned from known eukaryotic viral genomes.

Multilayer architecture

To ensure stringent control of false-positive classifications within the ML workflow, unmapped reads are backmapped to assembled contigs using *bwa-mem2* to obtain mapping statistics. Contigs supported by fewer than 100 mapped reads are excluded from downstream analysis. In addition, contigs are subjected to translated similarity search using *diamond* 2.1.16 (blastx mode)³⁴ against NCBI's clustered protein database (clustered_nr), enabling the detection of protein-level homology and further excluding spurious assemblies. The combination of coverage-based support, machine-learning-based viral prediction, and protein-level homology screening provides a multi-layered classification system designed to minimize false-positive signals while retaining high recall for divergent viral sequences.

Reporting

Detected viruses, together with all classification and filtering results, are summarized in a table that reports additional metrics including percent genome coverage and the proportion of high-quality read assignments (mapping quality ≥ 60). Viral abundances are visualized using a custom Python-based word cloud and an interactive pie chart generated with *KronaTools* 2.7.1³⁵. Coverage profiles are produced with *samtools coverage* using the `--histogram` flag²⁸. In addition, genome alignments and coverage profiles can be interactively inspected using *Integrative Genomics Viewer* 2.15.2 (IGV)³⁶. Highly divergent viral candidates are compiled in a separate table detailing contig length, *VirHunter* statistics, mapped read counts, and, when available, diamond-based taxonomic classification. Contig sequences of viral candidates can be copied and used for downstream analysis (e.g. manual BLAST search, phylogeny or wet-lab validation).

Quality-control metrics are summarized using *MultiQC* 1.34³⁷, including *FastQC* results (adapter content, sequence length distribution, per-sequence quality scores, mean quality scores, and sequence counts before and after trimming), *FastQ Screen* statistics (number and proportion of removed host and non-viral reads and remaining potential viral reads), and *Cutadapt* statistics (number and percentage of trimmed reads). All results are presented through the web interface and are additionally compiled into a downloadable HTML report. For local execution, a self-contained HTML report is generated using *snakemake's* reporting functionality. Visualizations of a sample report are shown in Figure 1B-1 to B-4.

Benchmarking strategy

To determine whether ViroNEXT maintains high recall while substantially reducing false-positive assignments, we evaluated its performance across a hierarchy of increasingly complex datasets. Synthetic benchmarks probed specific detection parameters under controlled conditions, experimental spike-in samples validated these findings under realistic laboratory conditions, and clinical specimens assessed applicability to naturally occurring infections.

Synthetic dataset

Synthetic datasets were generated to provide defined ground-truth samples with known viral composition, background and controlled sequence divergence. All synthetic reads were simulated *in silico* and processed through the complete ViroNEXT pipeline without parameter modification. Dataset generation procedures for each scenario are described below.

Viral coinfection scenario

Viral reads were simulated from the seven reference viruses of the "First WHO International Reference Panel for Adventitious Virus Detection by High-throughput Sequencing Technologies"¹³ (Supplementary Table 1). For each sample, 18 million paired-end reads were generated, corresponding to the theoretical sequencing depth of the spike-in benchmarking datasets used in this study. Reads were composed of 70% host sequences (human; substitution rate 0.2%), 20% of opportunistic bacterial and fungal sequences commonly found in skin (10% mutation rate each), and 10% viral reads. Specifically, each bacterial and fungal group comprised four species, with 450,000 reads simulated per species. The number of reads per virus varied between 30 and 1,000,000. In total, six samples with viral substitution rates of 0%, 5%, 10%, 15%, 20% and 25% were generated, deliberately spanning a broad dynamic range of viral abundances to challenge classification performance at different levels of viral representation.

Viral and background sequences were downloaded from NCBI using *efetch* from *Entrez direct* 16.2³⁸. *mutate.sh* from *BBTools* 39.17²⁴ was utilized to introduce substitution at a user-specified rate. Reads were simulated with the *art next-generation read simulator* 2016.06.05³⁹. Fragment length and standard deviation were obtained with the R package *BioanalyzeR* 0.10.1⁴⁰ and set to values representative for our spike-in data. Read profiles were generated using *art_profiler_illumina*³⁹ and representative fastq files from our sequencing runs to approximate realistic base qualities. Fastq files of the simulated reads from the different taxa were concatenated, compressed to gzipped format and

shuffled using `shuffle.sh` from the *BBTools* 39.1²⁴. Quality control was performed using *FastQC* 0.12.1²⁶. The entire workflow was implemented in a *snakemake* 9.16.3 pipeline and key parameters (e.g. number of reads per species, mutation rate, software version) were recorded in the report file for traceability purposes.

Environmental virome scenario

The synthetic environmental virome dataset was obtained from LaPierre *et al.* (2019)²² via the authors' GitHub repository and processed through the complete ViroNEXT pipeline using default parameters.

Diagnostic escape scenario

We developed a *snakemake* pipeline that generates synthetic viral genomes following predefined phylogenetic tree topologies. Branch length, expressed as expected substitutions per site, served as a proxy for evolutionary distance, enabling systematic simulation of viruses with increasing divergence from known references. Briefly, the pipeline retrieves related viral genomes based on user-provided accessions, constructs phylogenetic trees, inserts additional nodes at defined branch lengths representing divergent taxa (Figure 3a), generates the corresponding synthetic genomes, and simulates reads from these genomes using the procedure described above (default: 10,000 reads).

We chose to create divergent reads from five viruses that are frequently associated with humans (Hepatitis B, Torque teno virus, Human parvovirus B19, Human rhinovirus and Dengue virus) and combined those reads into one sample (Supplementary Figure 4). Sequence similarities for each divergent taxon are listed in Supplementary Table 2.

Sequences specified by the user were retrieved from NCBI using *efetch* from *Entrez direct* 16.2³⁸. Related taxa were determined using *blastn* search (*blast+* 2.7) against the *refseq_viral* database⁴¹ (access date: 2025-09-03). Their sequences were downloaded as mentioned before. Sequences were concatenated to a multifasta file and sequences were aligned using *MAFFT* 7.526 with `--auto` flag⁴². Maximum likelihood trees were constructed using *RAxML-NG* 1.2.2 with the General Time Reversible (GTR) model⁴³. Median branch length was inferred with *ete3* 3.1.3 toolkit⁴⁴ implemented in Python 3. Synthetic divergent taxa were added to the phylogenetic trees at the user-specified positions. Sequence dissimilarity was controlled by setting the branch length of the divergent taxa to

$$BL = BL_{multiplier} * BL_{median}$$

where BL is the resulting branch length of the divergent virus, BL_{multiplier} is a user-specified list of multipliers determining the range of sequence similarities (default = 0, 0.25, 0.50, 1.00, 2.00, 4.00 and 8.00) and BL_{median} is the median branch length of the tree.

A custom Python script was used to reformat data for divergent genome generation. Divergent genomes were produced as FASTA files using *Seq-Gen 1.3.4*⁴⁵ under the GTR substitution model. The generated sequences were then appended to the initial MAFFT alignment and subsequently re-aligned. Phylogenetic trees containing the divergent virus genomes were constructed as described earlier.

Reads derived from these mutated genomes and a background containing human, bacteria and fungi (as outlined above) were simulated, concatenated, shuffled, quality-controlled and processed through the ViroNEXT pipeline. After completion of the pipeline, predicted viral contigs were aligned to their corresponding synthetic genomes using *MAFFT* as mentioned before. Contigs were assigned to their source genomes based on sequence similarity metrics computed using a custom Python script (see Supplementary Figure 9 for an example).

Experimental spike-in datasets

Two independent spike-in datasets were analyzed to assess pipeline performance. The **WHO International Reference Panel dataset** originated from an evaluation of the "First WHO International Reference Panel for Adventitious Virus Detection by High-throughput Sequencing Technologies"³⁹ and was also used to assess limit of detection (LoD). The **multi-virus spike-in dataset** comprised hybridization-capture sequencing data generated as part of a study comparing shotgun and capture-based viral metagenomics with respect to breadth of viral detection (manuscript in preparation). In addition, the WHO spike-in dataset was further used to assess the ability of the machine-learning branch to detect novel viruses, defined here as viruses whose entire family was systematically excluded from both the reference database and the ML training data (see Novel virus scenario below). Shortly, to generate these datasets, the nucleic acids from 400 µL of the spiked samples were isolated using a KingFisher Duo Prime instrument (Thermo Fisher Scientific) with the MagMAX Viral/Pathogen Nucleic Acid Isolation Kit according to the manufacturer's instructions and eluted in 70 µL of elution buffer. cDNA and double-stranded DNA were generated from 40 µL of the eluate and used for the library preparation. The library was generated utilizing the Twist Library Preparation EF 2.0 kit with the Twist Universal Adapter System (Twist Bioscience) according to the manufacturer's instructions. A hybridization/capture step using the VirCapSeq-VERT probes was performed in pools of 12 libraries each, followed by a final PCR round and clean-up before sequencing. Paired-end sequencing was

performed in an Illumina Nextseq2000 instrument in a multiplex manner attempting 18 million reads per sample.

WHO International Reference Panel

These datasets were generated using viruses from the WHO International Reference Panel (Supplementary Table 1), which were combined into a viral mixture and spiked at concentrations ranging from 10^5 to 10^1 genome equivalents (GE)/mL per virus into either human plasma or Dulbecco's Modified Eagle Medium (DMEM) containing Adenovirus 5 at 10^9 GE/mL as background, representing a human plasma sample and a vector vaccine product, respectively. Three replicates were analyzed for each sample and concentration.

Multi-virus spike-in dataset

To assess ViroNEXT performance in a complex multi-virus sample, we analyzed a spike-in dataset of 21 viruses in human plasma background (Supplementary Table 7). The panel comprised seven viruses obtained from Paul Ehrlich Institute (PEI) and WHO international reference standards in lyophilized plasma and 14 viruses from in-house collection in cell culture media background. The viruses obtained from reference standards were diluted to 10^4 GE/mL, the in-house panel to 100 TCID₅₀/mL, with exception of Bovine enterovirus which was spiked-in at a concentration of 50 TCID₅₀/mL. Data from three replicates per sample were included in the analysis.

Novel virus scenario

The machine-learning branch was further evaluated using the WHO spike-in dataset in blood plasma (10^5 GE/mL). First, we generated seven modified versions of our custom-filtered RVDB database. For each version, all sequences belonging to one viral family represented in the WHO International Reference Panel were removed (NCBI taxon IDs: 39724, 2946187, 3044472, 11632, 11118, 11244 and 10780, corresponding to *Circoviridae*, *Spinareoviridae*, *Orthoherpesviridae*, *Retroviridae*, *Coronaviridae*, *Pneumoviridae* and *Parvoviridae*, respectively). Each modified database was used both to build the *bwa* reference index and to re-train the VirHunter model, ensuring that neither the reference-based nor the machine-learning branch had access to sequences from the excluded viral family. The resulting pipeline configurations were then evaluated using the WHO spike-in dataset.

All contigs classified as viral by the machine-learning branch were subsequently examined using *diamond* 2.1.16 in blastx mode and BLASTn. Contigs for which the subsequent similarity searches supported assignment to the corresponding viral family represented in the WHO International

Reference Panel were considered successfully recovered. These contigs were aligned to their corresponding reference genomes using *Minimap2* v2.3⁴⁶, and genome coverage was calculated with *samtools* coverage as described above.

Clinical scenario

Clinical performance was evaluated using sequencing data from two distinct sources comprising a total of 20 samples. The first set consisted of 13 metagenomic datasets selected from a previously published benchmarking study, representing diverse sample types from patients with encephalitis or viral respiratory infections⁴². The second set comprised 7 samples from an independent cohort of patients presenting with suspected arboviral infections (Zika, Dengue, or Chikungunya)¹⁹. Sample inclusion for the clinical evaluations was strictly constrained by the availability of sufficient nucleic acid remnants to perform orthogonal confirmatory PCR testing. This residual material was required to experimentally validate unexpected viral assignments generated by the pipelines. All datasets were processed through the complete ViroNEXT pipeline using default parameters, without manual curation or parameter modification.

Encephalitis and respiratory infections

Thirteen clinical metagenomic datasets from patients with encephalitis or viral respiratory infections were obtained from de Vries *et al.* (2021)¹⁸. The cohort included cerebrospinal fluid (CSF; n = 4), brain biopsies (n = 3), nasopharyngeal swabs (n = 3), nasal washings (n = 1), bronchoalveolar lavage (n = 1), and one plasma sample (n = 1). More details can be extracted from Supplementary Table 8 and de Vries *et al.* (2021)¹⁸.

Arbovirus cohort

We analyzed samples from the Zik01 cohort, a prospective, observational, longitudinal cohort conducted in Tapachula, Chiapas, Mexico, between 2016 and 2018, to determine the epidemiological and clinical presentations of patients with Zika, Chikungunya, or Dengue (ClinicalTrials.gov NCT02831699). This cohort enrolled participants presenting with an acute illness defined by fever and/or rash and with clinical suspicion of Zika, dengue, or chikungunya infection. Participants underwent longitudinal follow-up, and serum and urine samples were collected at predetermined study visits. Participants with negative RT-qPCR results for Zika virus, Dengue virus, and Chikungunya virus throughout the follow-up period were classified as having an acute illness of unknown origin (AIUO). For the present study, a subset of seven samples (four serum and three urine) from this sequenced AIUO group was selected to evaluate the performance and cross-pipeline concordance of ViroNEXT under real-world clinical conditions.

Performance evaluation

Pipeline performance was assessed at the species level using recall, precision, and F_1 score. Recall represents the proportion of expected viruses that were detected, precision the proportion of detected viruses that were expected, and the F_1 score the harmonic mean of recall and precision.

For spike-in and human samples, the origins of detected viruses that were not among the WHO panel were systematically investigated. They were neglected when calculating precision if they were likely to originate from blood plasma (e.g. pegiviruses and anelloviruses) or adenovirus background (human adenovirus) or if PCR confirmed their presence in the sample (see below).

The limit of detection (LOD) was defined per taxon as GE/mL of the most diluted spike-in sample with detectable virus.

All evaluations were conducted using fully predefined, objective inclusion criteria, without manual curation or subjective interpretation.

Statistical analysis

Statistical analyses were performed in R version 4.6.1 using glmmTMB (version 1.1.14), emmeans (version 2.0.4), and DHARMa (version 0.5.0). False-positive assignments were compared among pipelines using a negative-binomial generalized linear mixed-effects model (GLMM). False-positive counts were modeled as the response variable, with pipeline as a fixed effect and sample identity as a random intercept to account for repeated analysis of the same samples by multiple pipelines. The overall effect of pipeline was assessed using a likelihood-ratio test comparing models with and without the pipeline term. Pairwise comparisons were performed using estimated marginal means (emmeans) and expressed as rate ratios with 95% confidence intervals. Multiple comparisons were corrected using the Holm method.

Model adequacy was evaluated using simulated residual diagnostics implemented in the DHARMa package, including tests for dispersion, zero inflation and residual uniformity. No significant deviations from model assumptions were observed. As a sensitivity analysis, pipeline differences were additionally evaluated using Friedman's test followed by paired Wilcoxon signed-rank tests with Holm correction. All statistical tests were two-sided, and P values <0.05 were considered statistically significant; P values were adjusted for multiple testing where applicable.

Head-to-head comparison with existing pipelines

To compare the performance of ViroNEXT with existing pipelines, we screened the literature for previously benchmarked pipelines. de Vries *et al.* (2021)¹⁸ benchmarked 13 bioinformatic pipelines for metagenomic virus detection. From these, we included all pipelines in our comparison, that met the following criteria:

- maximum precision
- actively maintained
- accessible to non-bioinformaticians via a web interface

Following these inclusion criteria, Genome Detective (Pan-viral, version 2.21.4)¹⁴ and One Codex (version 7f6b028b9d964d87)¹⁵ were selected. Moreover, SeqtoID, (formerly Chan Zuckerberg ID (CZID, version 8.3.15) was included as its superior performance has been demonstrated in Kalantar *et al.* (2020)³. Reference databases could not be standardized across pipelines because each pipeline uses its own database resources. However, the database configuration was kept constant across all analyses within each pipeline. To compare pipeline performance without manual intervention, all pipelines were executed using default settings. Performance of the pipelines was compared using the performance metrics described above and the following evaluation criteria. For ViroNEXT, only assignments that passed all pipeline stages were considered, i.e., concordant assignments across the primary, secondary, and tertiary classification steps that also passed coverage filtering. Genome Detective reports results in two categories: assignments (strains with $\geq 50\%$ nucleotide identity) and discoveries (strains with low nucleotide identity and/or limited genome coverage). For comparative evaluation, only assignments were considered, as inclusion of discoveries resulted in unstable performance metrics and significantly increased false-positive rates in preliminary analyses. For One Codex, no read-count threshold was applied, as such thresholds systematically exclude low-abundance taxa, a finding confirmed in preliminary analyses. Evaluations were restricted to species level identifications, as inclusion of higher-level taxonomic assignments increased false-positive rates, particularly for CZID.

PCR validation

Validation by PCR was necessary to obtain robust precision metrics (Supplementary table 8). However, given the high number of viral assignments reported by some pipelines (sometimes exceeding 100 per sample) and the limited availability of sample material, we established predefined rules to prioritize confirmatory testing without compromising performance statistics. PCRs were carried out to confirm or reject the presence of a virus when exactly two pipelines detected the same taxon. In cases of high cross-pipeline concordance (three or more pipelines detected the same virus), no PCR was performed

and the virus was assumed to be a genuine component of the sample's virome. If a virus was detected by a single pipeline only, the assignment was considered a false-positive. Similarly, viruses from non-human hosts without zoonotic potential were considered false-positive. Viruses commonly associated with the sample background (e.g., anelloviruses in plasma samples) were not counted as false-positive assignments.

Databases

C-RVDB. Clustered Reference Viral Database (C-RVDB) version 29 was obtained from <https://rvdb.dbi.udel.edu/>. C-RVDB has been designed by the Center for Biologics Evaluation and Research (CBER) at the U.S. Food and Drug Administration (FDA) and is tailored to viral classification using NGS-generated sequencing data¹⁷. To optimize computational efficiency and reduce redundancy in this database, all sequences shorter than 1000 nt were removed from C-RVDB and the remaining sequences were then grouped by species. Subsequently, Mash distance between every pair of sequences within each group was computed⁴⁷. Sequences not derived from RefSeq that exhibited high similarity (Mash distance < 0.15) to at least one other sequence were removed. An overview of the distribution of sequences per species before and after filtering can be found in Supplementary Figure 10. Within the ViroNEXT workflow, *bwa-mem2*, *VirHunter* and *BLAST* analyses were based on the nucleotide version of C-RVDB¹⁷, while C-RVDB-prot version 29 was used for *Kaiju* classification⁴⁸.

diamond. A clustered, non-redundant protein database was obtained from <ftp://ftp.ncbi.nlm.nih.gov/blast/db/experimental/> (access date 2025-12-01) and processed for the use with *diamond* with *blastdbcmd* version 2.15.0+³² and *makedb*³⁴.

Abbreviations

CBER: Center for Biologics Evaluation and Research

C-RVDB: Clustered Reference Viral Database

de.NBI: German Network for Bioinformatics Infrastructure

FDA: Food and Drug Administration

GLMM: generalized linear mixed model

IQR : interquartile range

NGS: next-generation sequencing

SRA: NCBI Sequence Read Archive

Declarations

Ethics approval and consent to participate

The Zik01 study protocol was approved by the Research and Ethics Committees of the Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán (approval number INF-2636-18-19-1) and the Instituto Mexicano del Seguro Social (approval number 09-B5-61-2800/201600/2378). Study procedures were conducted in accordance with relevant guidelines and regulations across the participating clinical sites in Tapachula. Written informed consent was obtained from all participants.

Consent for publication

Not applicable.

Availability of data and materials

The synthetic benchmark datasets generated and analyzed in this study have been deposited in Zenodo under DOI: 10.5281/zenodo.20731382. The spike-in sequencing data from the WHO International Reference Panel are available in the NCBI Sequence Read Archive (SRA) under BioProject accession PRJNA1498577. The spike-in data of our big panel have been deposited under BioProject accession PRJNA1499427. Databases can be downloaded from <https://github.com/brama108/vironext> or requested directly with the authors.

ViroNEXT is freely available for research use as open-source software for local execution and through a publicly accessible web interface for cloud-based analysis. The complete workflow, source code, documentation and access to the web interface, is available at <https://github.com/brama108/vironext>. The pipeline version used for the analyses in this manuscript was v0.31, with the corresponding public release available as ViroNEXT v1.0.0. Data uploaded through the web interface, including intermediate files, are deleted after seven days.

Competing interests

The authors declare that they have no competing interests

Funding

The project was supported by the German Federal Ministry of Health (Bundesministerium für Gesundheit), based on a resolution of the German Bundestag, through the grant "BLOODVIR - Integrated Network for Monitoring and Identification of Infectious Disease Risks", funding code ZMI5-2521DAT400 to R.K. In addition, this work was supported by the de.NBI Cloud within the

German Network for Bioinformatics Infrastructure (de.NBI) and ELIXIR-DE (Forschungszentrum Jülich and W-de.NBI-001, W-de.NBI-004, W-de.NBI-008, W-de.NBI-010, W-de.NBI-013, W-de.NBI-014, W-de.NBI-016, W-de.NBI-022).

Authors' contributions

M.S.B. conceptualization, data curation, formal analysis, investigation, methodology, software, validation, writing-original draft and writing - review & editing. G.H.R.Q. conceptualization, formal analysis, investigation, methodology, project administration, resources, supervision, validation, writing-original draft and writing - review & editing. M.M. data curation, formal analysis, methodology, investigation, software, and validation. J.D. investigation, methodology, and validation. C.M. resources. X.-J.L. methodology. P.R. methodology, resources, and validation. M.H. formal analysis. P.D.S. methodology. L.E. project administration, writing - review & editing. J.W. resources. L.C. supervision, and methodology. G.M.R.-P. resources, supervision and writing – review & editing. W.I.L. resources, supervision and, writing – review & editing. J.B. and R.K. Conceptualization, funding acquisition, supervision, writing-original draft and writing – review & editing.

Acknowledgements

Not applicable.

.

(A)



(B-2)

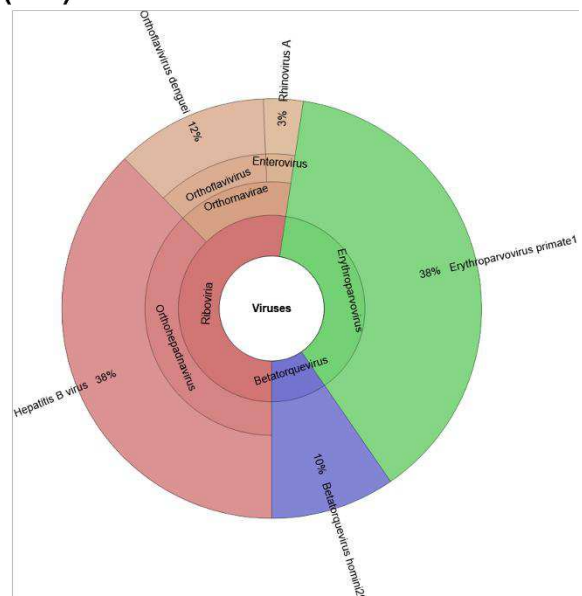


Figure 1 | ViroNEXT workflow and example report output. (A) Simplified overview of the ViroNEXT workflow for the detection of known viruses (black outlines) and highly divergent viral candidates (ML branch, blue outlines). Gray triangles schematically indicate the progressive reduction of

candidate viral assignments during downstream classification. (B) Representative elements of the interactive HTML report, including detected viral taxa and filtering results (B-1), coverage visualization and consensus validation (B-2, B-4) and abundance visualization (B-3).

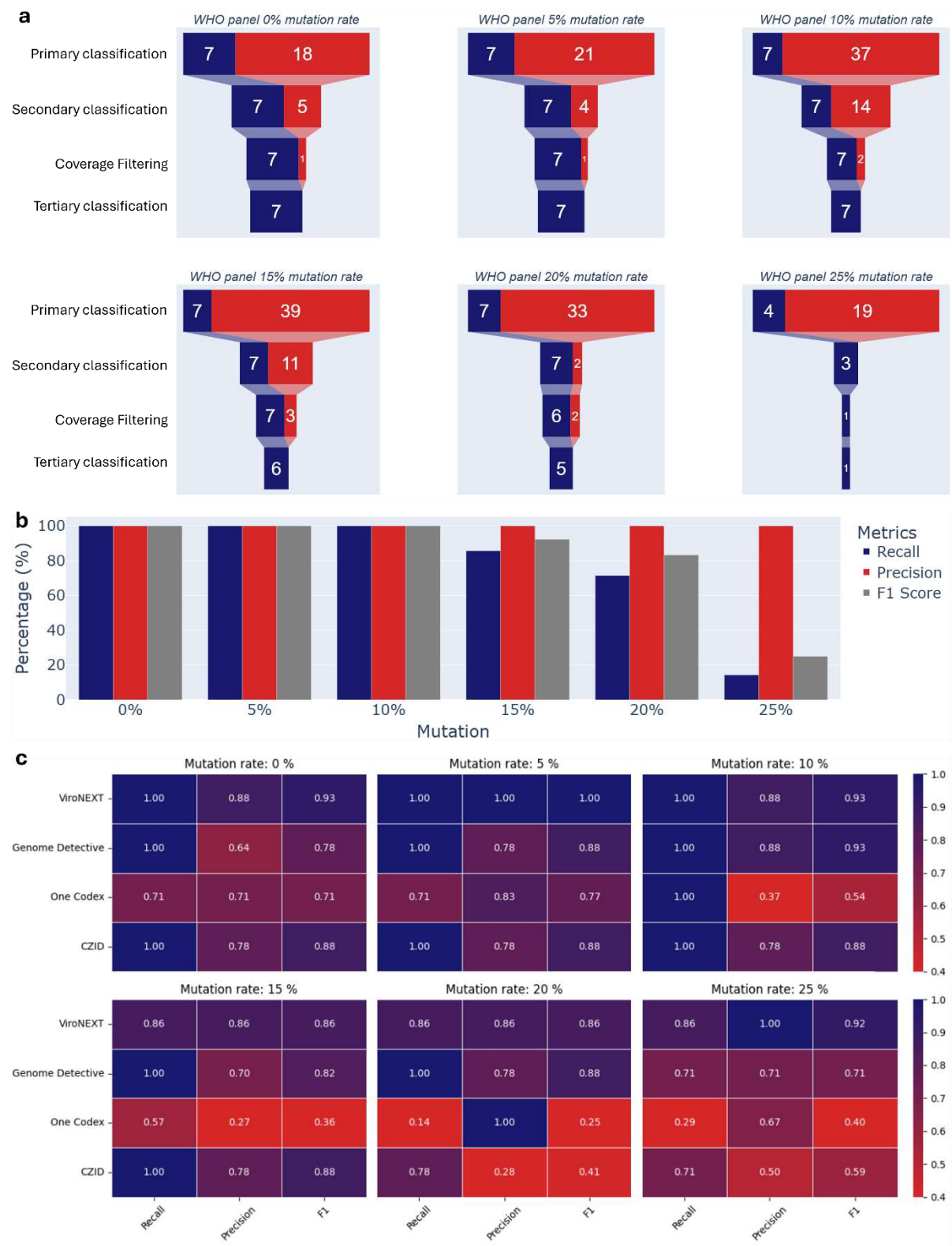


Figure 2 | Performance evaluation of ViroNEXT on synthetic WHO panel datasets with increasing sequence divergence.
a, Progression of true-positive (blue) and false-positive (red) viral assignments through the reference-based classification workflow of ViroNEXT across mutation rates ranging from 0% to 25%. Synthetic

datasets contained reads from seven viral species from the WHO viral panel. Each funnel plot shows the number of assignments remaining after sequential primary nucleotide classification of trimmed reads, secondary amino acid classification, coverage filtering, and tertiary consensus-sequence classification.

b, Overall performance of ViroNEXT (reference-based classification) across increasing mutation rates. Bar plots show recall, precision and F1 score for the combined classification workflow at mutation rates ranging from 0% to 25%.

c, Comparative performance of ViroNEXT, Genome Detective, One Codex, and CZID on synthetic WHO viral panel datasets with increasing mutation rates. Heatmaps show recall, precision, and F1 score across mutation rates from 0% to 25% combining reference- and ML-based classification

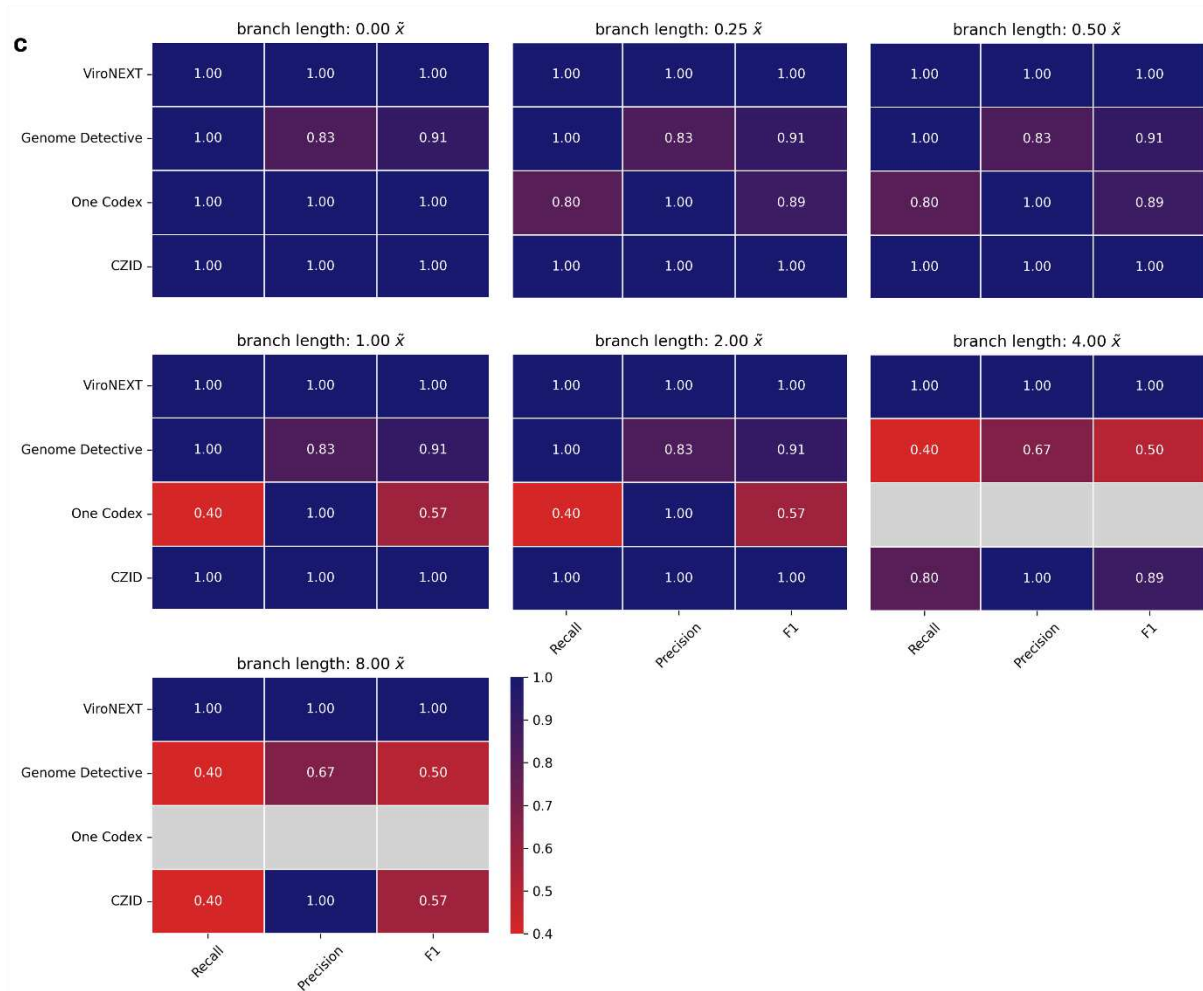
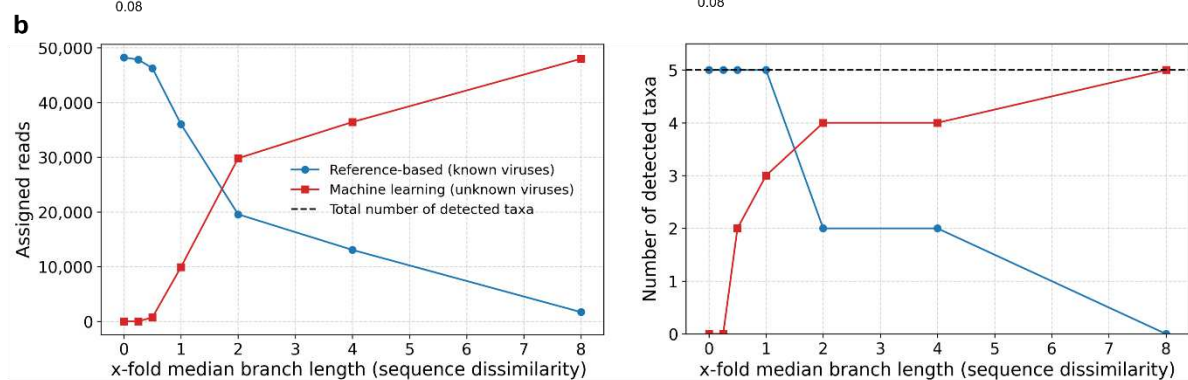


Figure 3 | Detection of evolutionarily divergent viral sequences by ViroNEXT. **a**, Representative phylogenies showing the original viral tree (left) and a synthetic tree generated by scaling branch lengths to four times the median branch length (right). The branch where the synthetic divergent taxon was introduced is highlighted in red, and the synthetic taxon is shown in blue. **b**, Complementary performance of the reference-based and machine-learning branches of ViroNEXT across increasing sequence divergence, summarized across all five viral phylogenies analyzed. Synthetic viral sequences were generated by scaling phylogenetic branch lengths relative to the median branch length of each viral phylogeny. Left, total number of assigned reads across all five simulated taxa by the reference-based (blue) and machine-learning (red) branches. Right, number of divergent taxa detected by each branch. The dashed line indicates the total number of simulated divergent taxa. **c**, Performance of ViroNEXT, Genome Detective, One Codex, and CZID across increasing sequence divergence. Heatmaps show recall, precision, and F_1 score at branch lengths ranging from 0× to 8× the median branch length. For ViroNEXT, values represent the combined output of the reference-based and machine-learning classification branches.

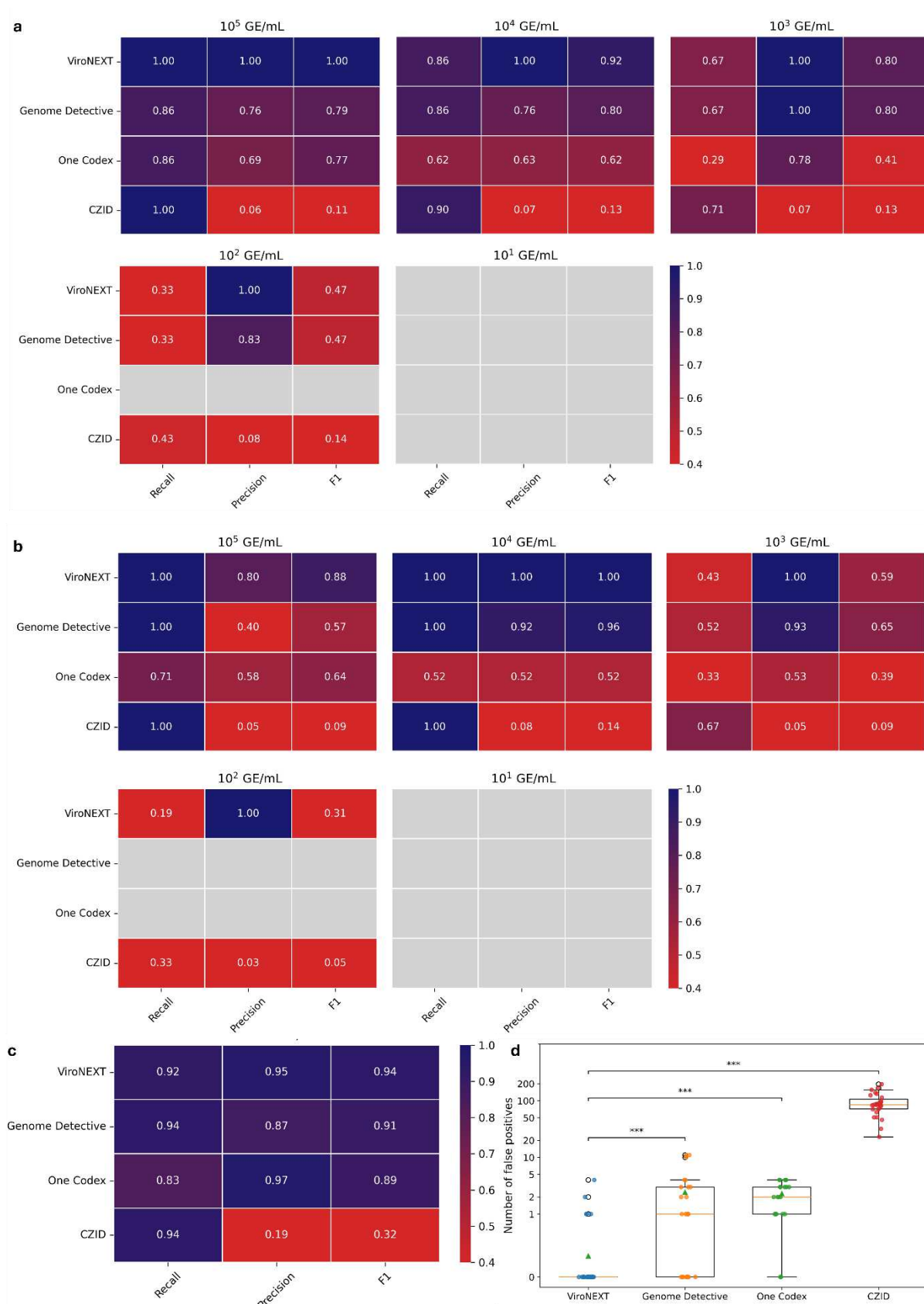
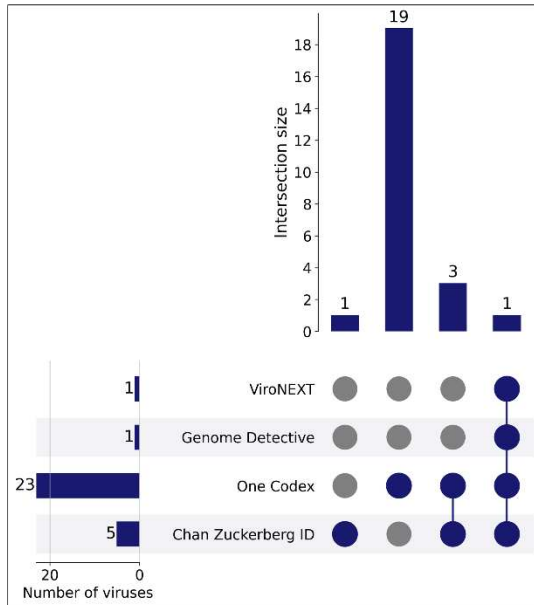


Figure 4 | Performance of ViroNEXT on viral spike-in datasets and large multi-virus panels. a, Performance of ViroNEXT and benchmarked pipelines on WHO viral panel datasets spiked into blood plasma across serial viral dilutions. **b,** Performance of ViroNEXT and comparator pipelines on WHO

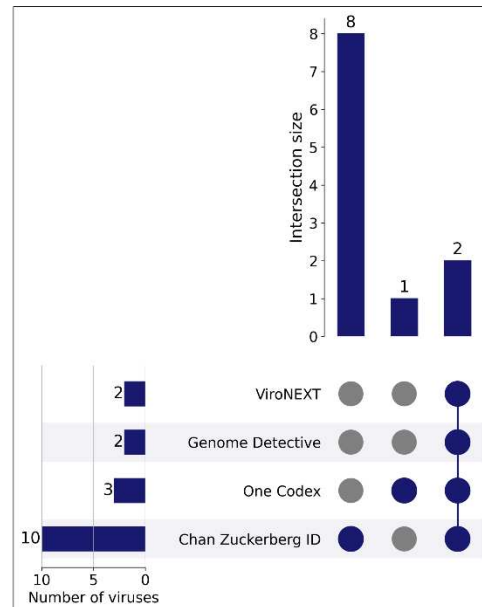
viral panel datasets spiked into an Adenovirus background. **c**, Comparative performance of ViroNEXT, Genome Detective, One Codex, and CZID on the large multi-virus panel comprising 21 viruses. **d**, False-positive assignments generated by four metagenomic classification pipelines across all spike-in experiments. Boxplots represent median (center line), interquartile range (IQR; box) and whiskers extending to $1.5 \times \text{IQR}$. Filled points represent individual sample replicates, open circles indicate values beyond the whiskers and green triangles indicate arithmetic means. The y axis is shown on a symmetric logarithmic (symlog) scale with a linear threshold of one false-positive assignment to visualize low counts while accommodating the broad range of values across pipelines. Differences between pipelines were assessed using a negative-binomial generalized linear mixed model (GLMM), followed by pairwise comparisons with Holm correction for multiple comparisons (ViroNEXT versus Genome Detective, $n = 24$ pairs; One Codex, $n = 21$ pairs; CZID, $n = 27$ pairs). Only samples with available results for both pipelines were included in each comparison. Asterisks indicate Holm-adjusted P values for comparisons with ViroNEXT ($***P_{\text{adj}} < 0.001$).

a

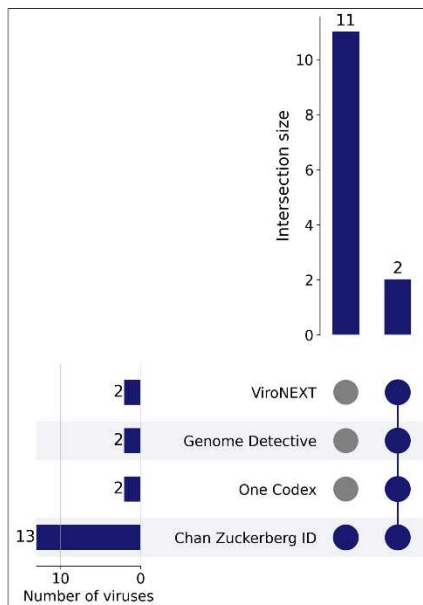
MXS_0549



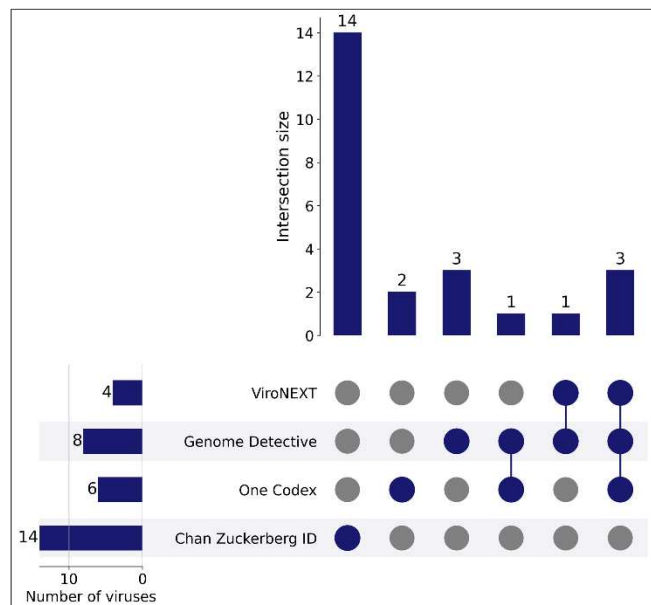
MXS_0567

**b**

MXU_0128



MXU_0218

**Figure 5 | Cross-pipeline concordance of viral assignments across clinical samples.**

UpSet plots summarizing the overlap of total viral species assignments (independent of the further classification as supported or unsupported) generated by ViroNEXT, Genome Detective, One Codex, and CZID across (a) serum and (b) urine samples of the Zik01 cohort. Horizontal bars indicate the total number of viral assignments per pipeline, whereas vertical bars represent the number of shared or unique assignments for each pipeline combination indicated by connected dots below.

Table 1 | Leave-one-out validation of the machine-learning branch using the WHO spike-in dataset. Viral families represented in the WHO International Reference Panel were removed from both the reference database and the machine-learning training dataset before rerunning the ViroNEXT pipeline. Genome coverage achieved by the machine-learning branch was compared with that obtained by the reference-based branch. In the reference-based column, values outside the brackets indicate genome coverage after removal of the corresponding viral family from the reference database, whereas values in brackets indicate genome coverage obtained with the unmodified reference database. Values represent mean genome coverage (%) $\pm$ standard deviation.

Species	Family	Genome coverage (%) ¹	
		Machine-learning	Reference-based Family removed (standard)
<i>Lymphocryptovirus humangamma4</i>	<i>Orthoherpesviridae</i>	66.5 $\pm$ 5.5	0 (91.1 $\pm$ 0.9)
<i>Feline leukemia virus</i>	<i>Retroviridae</i>	68.6 $\pm$ 18.1	0 (97.2 $\pm$ 3.0)
<i>Betacoronavirus 1</i> *	<i>Coronaviridae</i>	0 $\pm$ 0.0	0 (7.0 $\pm$ 5.0)
<i>Protoparvovirus rodent1</i>	<i>Parvoviridae</i>	91.7 $\pm$ 6.6	0 (98.2 $\pm$ 1.6)
<i>Circovirus porcine1</i>	<i>Circoviridae</i>	100.0 $\pm$ 0.0	0 (100.0 $\pm$ 0.0)
<i>Mammalian orthoreovirus 1</i>	<i>Spinareoviridae</i>	91.3 $\pm$ 10.4	0 (100.0 $\pm$ 0.0)
<i>Orthopneumovirus hominis</i>	<i>Pneumoviridae</i>	27.4 $\pm$ 15.0	0 (80.6 $\pm$ 16.2)

¹ Genome coverage values refer to the respective viral representatives within each family included in the WHO panel.

* *Betacoronavirus 1* was also not detected by PCR at concentration lower than 10⁶ GE/mL

References

1. Carbo, E. C. *et al.* Coronavirus discovery by metagenomic sequencing: a tool for pandemic preparedness. *Journal of clinical virology : the official publication of the Pan American Society for Clinical Virology* **131**, 104594; 10.1016/j.jcv.2020.104594 (2020).
2. Choi, C.-H. *et al.* Monkeypox Virus Genomic Analysis in the Republic of Korea: A Comparison of Metagenomic- and Probe Hybridization Capture Sequencing Methods. *Public Health Weekly Report* **17**, 859–873; 10.56786/PHWR.2024.17.20.3 (2024).
3. Kalantar, K. L. *et al.* IDseq-An open source cloud-based pipeline and analysis service for metagenomic pathogen detection and monitoring. *GigaScience* **9**; 10.1093/gigascience/giaa111 (2020).
4. López-Labrador, F. X. *et al.* Recommendations for the introduction of metagenomic high-throughput sequencing in clinical virology, part I: Wet lab procedure. *Journal of clinical virology : the official publication of the Pan American Society for Clinical Virology* **134**, 104691; 10.1016/j.jcv.2020.104691 (2021).

5. Sukhorukov, G. *et al.* VirHunter: A Deep Learning-Based Method for Detection of Novel RNA Viruses in Plant Sequencing Data. *Frontiers in bioinformatics* **2**, 867111; 10.3389/fbinf.2022.867111 (2022).
6. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265–269; 10.1038/s41586-020-2008-3 (2020).
7. Feng, H., Shuda, M., Chang, Y. & Moore, P. S. Clonal integration of a polyomavirus in human Merkel cell carcinoma. *Science (New York, N.Y.)* **319**, 1096–1100; 10.1126/science.1152586 (2008).
8. Finkbeiner, S. R. *et al.* Identification of a novel astrovirus (astrovirus VA1) associated with an outbreak of acute gastroenteritis. *Journal of virology* **83**, 10836–10839; 10.1128/JVI.00998-09 (2009).
9. Briese, T. *et al.* Genetic detection and characterization of Lujo virus, a new hemorrhagic fever-associated arenavirus from southern Africa. *PLoS pathogens* **5**, e1000455; 10.1371/journal.ppat.1000455 (2009).
10. Pérot, P., Bigot, T., Temmam, S., Regnault, B. & Eloït, M. Microseek: A Protein-Based Metagenomic Pipeline for Virus Diagnostic and Discovery. *Viruses* **14**; 10.3390/v14091990 (2022).
11. Charlebois, R. L., Ng, S. H. S., Gisonni-Lex, L. & Mallet, L. Cataloguing the taxonomic origins of sequences from a heterogeneous sample using phylogenomics: applications in adventitious agent detection. *PDA journal of pharmaceutical science and technology* **68**, 602–618; 10.5731/pdajpst.2014.01023 (2014).
12. Buddle, S. *et al.* Evaluating metagenomics and targeted approaches for diagnosis and surveillance of viruses. *Genome medicine* **16**, 111; 10.1186/s13073-024-01380-x (2024).
13. Khan, A. S., Chin, P.-J., Fuentes, S. M., Tsou, J.-H., Study Group. A collaborative study to evaluate the proposed first WHO International Reference Panel for adventitious virus detection in biological products by high-throughput sequencing (HTS) technologies. *WHO Expert Committee on Biological Standardization (WHO/BS/2024.2471 (World Health Organization, 2024))*.
14. Vilsker, M. *et al.* Genome Detective: an automated system for virus identification from high-throughput sequencing data. *Bioinformatics (Oxford, England)* **35**, 871–873; 10.1093/bioinformatics/bty695 (2019).
15. Minot, S. S., Krumm, N. & Greenfield, N. B. *One Codex: A Sensitive and Accurate Data Platform for Genomic Microbial Identification* (2015).
16. LaPierre, N. *et al.* MiCoP: microbial community profiling method for detecting viral and fungal organisms in metagenomic samples. *BMC genomics* **20**, 423; 10.1186/s12864-019-5699-9 (2019).
17. Goodacre, N., Aljanahi, A., Nandakumar, S., Mikailov, M. & Khan, A. S. A Reference Viral Database (RVDB) To Enhance Bioinformatics Analysis of High-Throughput Sequencing for Novel Virus Detection. *mSphere* **3**; 10.1128/mSphereDirect.00069-18 (2018).
18. Vries, J. J. C. de *et al.* Benchmark of thirteen bioinformatic pipelines for metagenomic virus diagnostics using datasets from clinical samples. *Journal of clinical virology : the official publication of the Pan American Society for Clinical Virology* **141**, 104908; 10.1016/j.jcv.2021.104908 (2021).
19. Belaunzarán-Zamudio, P. F. *et al.* Different epidemiological profiles in patients with Zika and dengue infection in Tapachula, Chiapas in Mexico (2016-2018): an observational, prospective cohort study. *BMC infectious diseases* **21**, 881; 10.1186/s12879-021-06520-x (2021).
20. Ren, J. *et al.* Identifying viruses from metagenomic data using deep learning. *Quantitative biology (Beijing, China)* **8**, 64–77; 10.1007/s40484-019-0187-4 (2020).
21. Mölder, F. *et al.* Sustainable data analysis with Snakemake. *F1000Res* **10**, 33; 10.12688/f1000research.29032.1 (2021).
22. ValidateFastq. <https://biopet.github.io/validatefastq/0.1.1/index.html> (2018).

23. Wingett, S. W. & Andrews, S. FastQ Screen: A tool for multi-genome mapping and quality control. *F1000Res* **7**, 1338; 10.12688/f1000research.15931.2 (2018).
24. Bushnell, B. BBMap: A Fast, Accurate, Splice-Aware Aligner. Available at <https://github.com/bbushnell/BBTools> (2014).
25. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet j.* **17**, 10; 10.14806/ej.17.1.200 (2011).
26. Andrews, S. FastQC: A Quality Control Tool for High Throughput Sequence Data (2010).
27. Md, V., Misra, S., Li, H. & Aluru, S. Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. *IEEE Parallel and Distributed Processing Symposium (IPDPS)* (2019).
28. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* **25**, 2078–2079; 10.1093/bioinformatics/btp352 (2009).
29. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics (Oxford, England)* **31**, 2032–2034; 10.1093/bioinformatics/btv098 (2015).
30. Menzel, P., Ng, K. L. & Krogh, A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nature communications* **7**, 11257; 10.1038/ncomms11257 (2016).
31. Vries, J. J. C. de *et al.* Recommendations for the introduction of metagenomic next-generation sequencing in clinical virology, part II: bioinformatic analysis and reporting. *Journal of clinical virology : the official publication of the Pan American Society for Clinical Virology* **138**, 104812; 10.1016/j.jcv.2021.104812 (2021).
32. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC bioinformatics* **10**, 421; 10.1186/1471-2105-10-421 (2009).
33. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology : a journal of computational molecular cell biology* **19**, 455–477; 10.1089/cmb.2012.0021 (2012).
34. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nature methods* **12**, 59–60; 10.1038/nmeth.3176 (2015).
35. Ondov, B. D., Bergman, N. H. & Phillippy, A. M. Interactive metagenomic visualization in a Web browser. *BMC bioinformatics* **12**, 385; 10.1186/1471-2105-12-385 (2011).
36. Robinson, J. T. *et al.* Integrative genomics viewer. *Nature biotechnology* **29**, 24–26; 10.1038/nbt.1754 (2011).
37. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics (Oxford, England)* **32**, 3047–3048; 10.1093/bioinformatics/btw354 (2016).
38. Kans, J. *Entrez® Direct: E-utilities on the Unix Command Line. 2013 Apr 23 [Updated 2025 Mar 25]. In: Entrez® Programming Utilities* (National Center for Biotechnology Information, Bethesda, 2010).
39. Huang, W., Li, L., Myers, J. R. & Marth, G. T. ART: a next-generation sequencing read simulator. *Bioinformatics (Oxford, England)* **28**, 593–594; 10.1093/bioinformatics/btr708 (2012).
40. Foley, J. bioanalyzeR. Available at <https://github.com/jwfoley/bioanalyzeR> (2023).
41. Brister, J. R., Ako-Adjei, D., Bao, Y. & Blinkova, O. NCBI viral genomes resource. *Nucleic acids research* **43**, D571–7; 10.1093/nar/gku1207 (2015).
42. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution* **30**, 772–780; 10.1093/molbev/mst010 (2013).
43. Kozlov, A. M., Darriba, D., Flouri, T., Morel, B. & Stamatakis, A. RAXML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics (Oxford, England)* **35**, 4453–4455; 10.1093/bioinformatics/btz305 (2019).

44. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Molecular biology and evolution* **33**, 1635–1638; 10.1093/molbev/msw046 (2016).
45. Rambaut, A. & Grassly, N. C. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Computer applications in the biosciences : CABIOS* **13**, 235–238; 10.1093/bioinformatics/13.3.235 (1997).
46. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics (Oxford, England)* **34**, 3094–3100; 10.1093/bioinformatics/bty191 (2018).
47. Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome biology* **17**, 132; 10.1186/s13059-016-0997-x (2016).
48. Bigot, T., Temmam, S., Pérot, P. & Eloit, M. RVDB-prot, a reference viral protein database and its HMM profiles. *F1000Res* **8**, 530; 10.12688/f1000research.18776.2 (2019).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Braunetalsupplfigurestable.docx](#)