PathoLive—Real-Time Pathogen Identification from Metagenomic Illumina Datasets

Simon H. Tausch [1,2,3,*,†], Tobias P. Loka [2,4,†], Jakob M. Schulze [2], Andreas Andrusch [2,3], Jeanette Klenner [3], Piotr Wojciech Dabrowski [2,5], Martin S. Lindner [2], Andreas Nitsche [3] and Bernhard Y. Renard [2,4]

1   National Study Centre for Sequencing in Risk Assessment, Department Biological Safety, German Federal Institute for Risk Assessment, 10589 Berlin, Germany
2   Bioinformatics Division (MF 1), Department for Methods Development and Research Infrastructure, Robert Koch Institute, 13353 Berlin, Germany
3   Centre for Biological Threats and Special Pathogens, Highly Pathogenic Viruses (ZBS 1), 13353 Berlin, Germany
4   Digital Engineering Faculty, Hasso Plattner Institute, University of Potsdam, 14482 Potsdam, Germany
5   School of Computing, Communication and Business (Faculty 4), HTW Berlin—University of Applied Sciences, 12459 Berlin, Germany
*   Correspondence: simon.tausch@bfr.bund.de
†   These authors contributed equally to this work.

Abstract: Over the past years, NGS has become a crucial workhorse for open-view pathogen diagnostics. Yet, long turnaround times result from using massively parallel high-throughput technologies as the analysis can only be performed after sequencing has finished. The interpretation of results can further be challenged by contaminations, clinically irrelevant sequences, and the sheer amount and complexity of the data. We implemented PathoLive, a real-time diagnostics pipeline for the detection of pathogens from clinical samples hours before sequencing has finished. Based on real-time alignment with HiLive2, mappings are scored with respect to common contaminations, low-entropy areas, and sequences of widespread, non-pathogenic organisms. The results are visualized using an interactive taxonomic tree that provides an easily interpretable overview of the relevance of hits. For a human plasma sample that was spiked in vitro with six pathogenic viruses, all agents were clearly detected after only 40 of 200 sequencing cycles. For a real-world sample from Sudan, the results correctly indicated the presence of Crimean-Congo hemorrhagic fever virus. In a second real-world dataset from the 2019 SARS-CoV-2 outbreak in Wuhan, we found the presence of a SARS coronavirus as the most relevant hit without the novel virus reference genome being included in the database. For all samples, clinically irrelevant hits were correctly de-emphasized. Our approach is valuable to obtain fast and accurate NGS-based pathogen identifications and correctly prioritize and visualize them based on their clinical significance: PathoLive is open source and available on GitLab and BioConda.

Keywords: NGS; metagenomics; viruses; infectious diseases; diagnostics; live sequencing

1. Introduction

The identification of pathogens directly from patient samples is a major clinical need. While highly accurate pathogen detection methods such as polymerase chain reaction (PCR), cell culture, or amplicon sequencing exist, such routine procedures often fail to identify the underlying cause of a patient's symptoms due to their targeted behavior [1–4]. As a complementary approach, metagenomics next-generation sequencing (NGS) has been proposed as a valuable technique for clinical application. NGS facilitates the detection and characterization of pathogens without a priori knowledge about candidate species. Further, it generates a sufficient amount of data to detect even lowly abundant pathogens without targeted amplification of specified sequences allowing for hypothesis-free diagnostic analysis.

Current tools to address NGS-based pathogen identification can be divided into two major categories, either aiming to discover yet unknown genomes [5–15] or to detect known organisms in a sample [16–32]. From an algorithmic perspective, a further distinction can be made between alignment-based methods, alignment-free methods, or combinations of both. While alignment-free methods usually deliver faster results, alignment-based methods potentially allow for a more extensive characterization of the sample.

Regardless of the algorithmic approach, existing methods based on unbiased metagenomics NGS face various obstacles, especially concerning the ranking of the results according to their clinical relevance and the long overall turnaround time [33–40]. The lack of good ranking methods is based on the fact that the distinction between clinically relevant and irrelevant data is not trivial. First, the dominating part of the sequences in a patient sample usually originates from the host genome. Second, there are nucleic acids of various species that are usually of low clinical relevance such as endogenous retroviruses (ERV) or non-pathogenic bacteria which commonly colonize a person. For these reasons, the number of reads hinting towards a relevant pathogen can be as low as a handful of individual reads. To put it more generally, it is a widespread misconception to rely only on quantitative measures when ranking the importance of candidate hits as not the amount but the uncommonness of a species in a given sample may give critical indications on its relevance. Based on the premise that a large proportion of the produced reads may stem from the host genome, species irrelevant for diagnosis, or common contaminations, even highly accurate methods struggle with false-positive hits potentially concealing the relevant results. This central problem is getting worse when considering that even microbial databases are contaminated with human sequences [41]. Existing pipelines tackle this problem in different ways. One common strategy is to ignore sequences that occur in a reference database of host and contaminating sequences [9,10,18,26,28,30]. While facilitating cleaner results, this approach may lead to a premature rejection of relevant sequences and does not solve the problem of human contaminations in reference databases as those "derive primarily from high-copy human repeat regions, which themselves are not adequately represented in the current human reference genome" [41]. Further, the definition of precise contamination databases proves rather difficult and has not yet been adequately solved. Thus, deleting any results to gain a better overview comes at great risk of overlooking the true cause of an infection. A different strategy is intensity filters, as implemented, e.g., in SLIMM [19], that disregard sequences with low genome coverages. As the author states, this step eliminates many genomes which introduces the risk of losing information that might be relevant in the following diagnostic process. This problem even intensifies for marker-gene-based methods such as MetaPhlAn2 [23], as large parts of the sequenced reads cannot be assigned due to the miniaturized reference database. While this may lead to a better ratio of seemingly relevant assigned reads to those from the background, it comes with the risk of disregarding relevant candidates.

Another fundamental problem of NGS-based pathogen identification approaches is the fact that sequencing and analysis are very time consuming. Even when considering the reduction in sequencing time in the last years, current mid- and high-throughput devices still have maximum runtimes of more than a day (NextSeq 550) and up to two (NovaSeq 6000) or three days (HiSeq X), respectively. The resulting turnaround times of two to four days including data processing and analysis are not short enough for many critical scenarios such as sepsis and infectious disease outbreaks. To obtain actionable results within an appropriate time frame it is crucial to reduce the time span from sample receipt to diagnosis. However, existing approaches to speed up NGS-based diagnostics come with significant disadvantages such as a highly reduced throughput and data quality [42], massive reduction of analyzed reads or targets [43] or the need for specialized hardware that involves additional costs and relatively low flexibility to adapt the workflow to a given scenario [44]. An actual approach for taxonomic classification of NGS data during runtime of sequencing is implemented in LiveKraken, a real-time version of the well-known Kraken software [45]. However, by not providing positional information in the

results, a sequence-based ranking to determine the relevance of hits is not possible with this approach.

As a general complement to real-time analysis of short-read sequencing data, there are several promising studies for pathogen detection using the MinION handheld device which is particularly useful for field studies and produces longer reads of up to several hundred kilobase pairs. While allowing very fast throughput times, these devices yield only approximately a million reads with comparably low per-base qualities, limiting their areas of application to targeted sequencing so far [5,42,46–48].

The currently high turnaround times from sample arrival to final diagnosis make it necessary to develop efficient methods to generate, analyze, and understand large metagenomics datasets in an accurate and quick manner to pave the way for NGS as a standard tool for clinical diagnostics. This enforces NGS-based diagnostics workflows to generate and evaluate large numbers of reads to facilitate adequate sequencing depths while reducing the time span between sample receipt and diagnosis. To overcome the named obstacles, we present PathoLive, an NGS-based real-time pathogen detection tool. We present an innovative approach to handle the occurrence of common contaminations, background data, and irrelevant species in a single step. To tackle the problem of long overall turnaround times, we based our novel approach on the real-time read mapper HiLive2 which enables the analysis of sequencing data while an Illumina sequencer is still running [49]. This enables PathoLive to perform nucleotide-level analysis based on NGS providing an open view and high accuracy in short turnaround times while generating an intuitive and interactive visualization of results that highlights organisms of high clinical significance.

2. Methods

2.1. Implementation

Our workflow follows a different paradigm than other frameworks to tackle the existing problems, as shown in Figure 1: (i) prepare informative, well-defined reference databases, (ii) automatically define contaminating or non-pathogenic sequences beforehand, (iii) use HiLive2 for accurate real-time alignment of Illumina sequencing data, (iv) visualize the potential risk of candidate pathogens and present results in an intuitive, comprehensible manner. The details on the modules for each of these steps are provided in the following paragraphs:

(i) Preparation of reference databases: In order to save computational effort during the analysis, reference databases including the full taxonomic lineage of organisms are prepared before the first execution of PathoLive. For this purpose user selectable databases, for example, the RefSeq Genomic Database [50], are downloaded from the File Transfer Protocol (FTP) servers of the National Center for Biotechnology Information (NCBI) and annotated accordingly with taxonomic information from the NCBI Taxonomy Database. While preserving the original NCBI annotation of each sequence, additional information is appended to the sequence header. This information consists of each taxonomic identifier (TaxID), rank, and name of each taxon in the lineage of an organism. Afterwards, user-definable sub-databases of taxonomic clades relevant for a distinct pathogen search are automatically created. For the experiments in this manuscript, we focused on viruses. The database updater used for this purpose is available at https://gitlab.com/rki_bioinformatics/database-updater (accessed on 23 August 2022). The viral database used in this manuscript can be downloaded as a single compressed FASTA file from Zenodo (https://doi.org/10.5281/zenodo.2536788, accessed on 23 August 2022) and is ready to use for viral diagnostics with PathoLive.
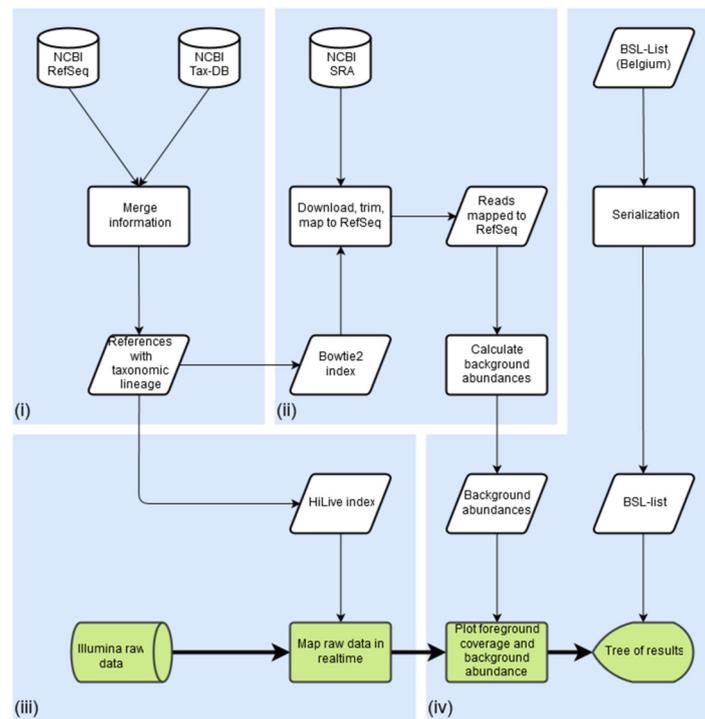
Figure 1. Workflow of PathoLive including four main modules. (**i**) Automated download and taxonomic tagging of reference information from NCBI RefSeq; (**ii**) NGS datasets from the 1000 Genomes Project are downloaded, trimmed, and searched for database sequences from step (**i**), marking abundant stretches as clinically irrelevant; (**iii**) reads from the clinical sample are mapped in real time, producing intermediate alignment files; (**iv**) results are visualized in an easily understandable manner, providing all available information while pointing to the most relevant results. Only the steps highlighted in green are calculated in execution time, steps in white are precomputation. Graphical results are presented only minutes after the sequencer finishes a cycle if desired.

(**ii**) **Identification and labeling of clinically irrelevant hits**: A main obstacle in NGS-based diagnostics is the large amount of background noise contained in the data. This includes various sources of contamination such as sequencing artifacts, ambiguous references, and clinically irrelevant species, which hinder a quick evaluation of a dataset. Defining an exhaustive set of possible contaminations is a yet unachieved goal. Furthermore, deleting such sequences carries the risk of losing relevant results. Since in this step raw sequencing data from a human host is examined, the logical conclusion is to contrast it to comparable raw datasets instead of processed genomes. Instead of deleting the background and risking the loss of relevant information, we implemented a method to define and mark all kinds of undesired signals on the basis of comparable datasets from freely available resources. For this purpose, raw data from 236 randomly selected datasets from the 1000 Genomes Project Phase 3 [51] were downloaded, assuming that a large majority of the participants in the 1000 Genomes Project were not acutely ill with an infectious disease. The full list of selected datasets is provided in the supplementary material (Section S3.3). The reads are quality trimmed using Trimmomatic [52] and mapped to the selected pathogen reference database using Bowtie2 [53]. Whenever a stretch of a sequence is covered once or more in a dataset from the 1000 Genomes Project, the overall background coverage of these bases is increased by one. Coverage maps of all references from the pathogen database are stored in the serialized pickle file format. Stretches of DNA found in this data are marked as of lower clinical significance and visualized as such in later steps of the workflow. The coverage maps of the background abundances are plotted in red color against the coverage maps of the reads from the patient dataset in green color on the same reference (Figure 2). This enables highlighting presumably relevant results

without discarding other candidate pathogens, giving the researcher the best options to interpret the results in-depth but still in an efficient manner. The code for the generation of these databases is part of PathoLive.
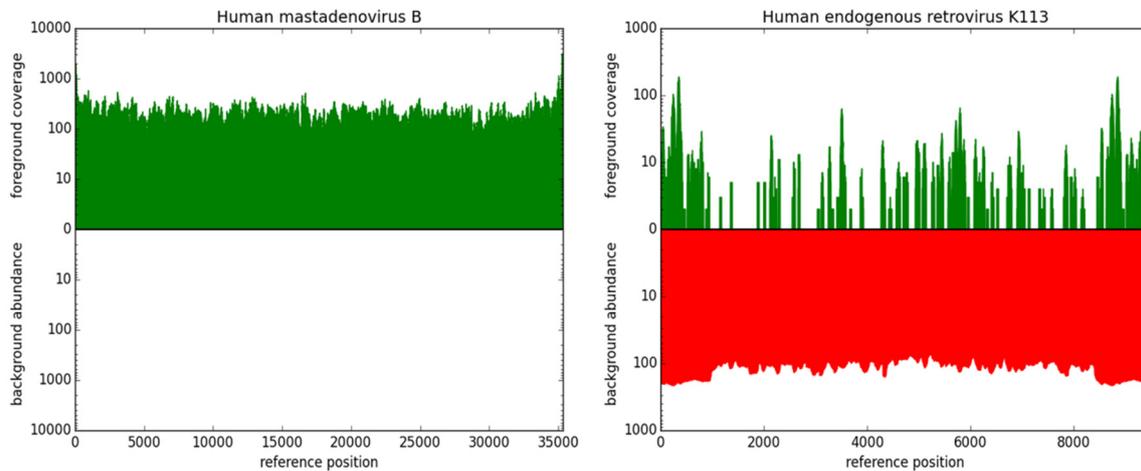

Figure 2. Two examples of fore- and background coverage plots. The upper, green bars show the coverage of a given genome in the foreground dataset, namely the reads sequenced from the patient sample. The lower, red part indicates in how many datasets from the 1000 Genomes Project a sequence is abundant. Bases covered in background datasets are regarded as less informative. Left: Fully covered genome of human mastadenovirus B, showing no hits resulting from data from the 1000 Genomes Project. Right: Coverage of human endogenous retrovirus (HERV) K113, partly covered in the patient dataset and completely covered in ~110 datasets from the 1000 Genomes Project. Based on these illustrations, human mastadenovirus B can be considered a relevant hit while HERV K113 is rightly found in the dataset, but not considered a clinically relevant candidate due to its common prevalence in healthy human individuals.

(iii) Using HiLive2 for real-time alignment of reads: We used HiLive2 (version 2.1) to produce real-time alignments of intermediate sequencing results. Thereby, the raw sequencing data is directly loaded in raw BCL file format without the need to perform a file conversion step. Alignments are updated with each new sequencing cycle and output in BAM format can be created for any sequencing cycle. As changes in the mapping positions mainly occur in early sequencing cycles, we recommend creating output in shorter intervals at the beginning of sequencing. Options for integrated demultiplexing and adapter trimming are available. For algorithmic details of HiLive2, we refer to Loka, Tausch and Renard [49].

(iv) Visualization and hazardousness classification: A key hurdle in a rapid diagnostics workflow, which is often underestimated, is the presentation of results in an intuitive way. Many promising efforts have been made by different tools, e.g., providing coverage plots [30,54] or interactive taxonomy explorers [12,28]. While being hard to measure and thus often ignored, the time it takes for groups of experts to assess the results and come to a correct conclusion should be considered. Our browser-based, interactive visualization is implemented in JavaScript using the data visualization library D3 [55]. For an example of the visualization, see Figure 3. While providing all available information on demand, the structure of a taxonomic tree allows an intuitive overview. Detailed measures are available on the genus, family, species, and sequence level. For the calculation of scores for a given node n, we define $t(n)$ as the total number of read alignments to an underlying species of n. $b(n)$ is the total number of bases being covered by all reads with respect to n. Accordingly, $b_{bg}(n)$ describes the number of bases being covered by the background database and $b_{fg \setminus bg}(n)$ is the number of bases being covered by the foreground but not by the background data. In total, we provide three different scores for each node n of the tree:

(a) Total hits T_n, representing the total number of hits to all underlying sequences in this branch: $T_n := t(n)$, representing the total abundance of a clade.

(b) Unambiguous bases U_n, representing the total number of bases covered in the foreground data but not in any background dataset: $U_n := b_{fg \setminus bg}(n)$

(c) Weighted score W_n, being the ratio of unambiguous bases for the foreground data to the number of bases covered by the background database and logarithmically weighted by the total number of alignments: $W_n := \frac{U_n}{max(b_{bg}(n),\,1)} \cdot \log(T_n)$
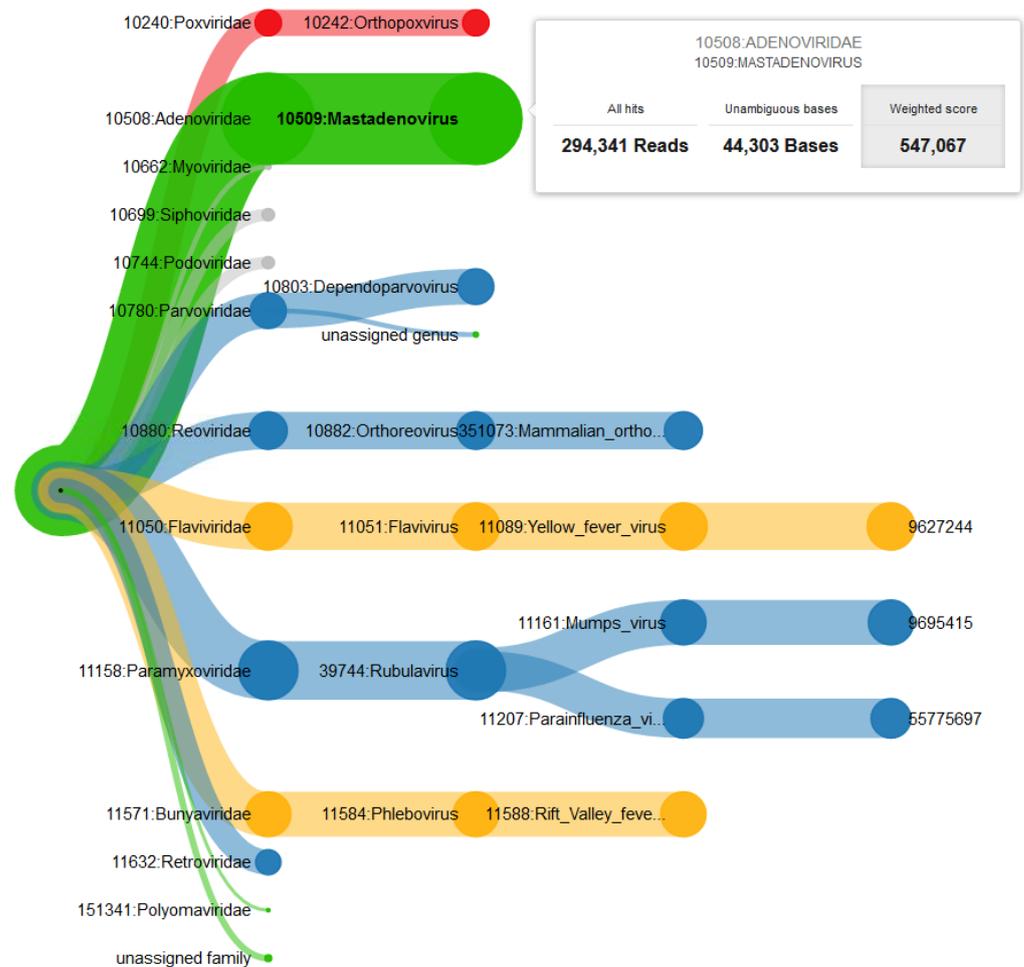

Figure 3. Example of the interactive taxonomic tree of results. The results show the described plasma sample at cycle 80 based on the weighted score. Thickness of the branches denotes the sum of scores of underlying sequences. The color codes for the maximum of the underlying BLS-levels (red = 4, yellow = 3, blue = 2, green = 1 or undefined; phages are shown in grey). On mouse-over, detailed information (here on genus Mastadenovirus) is displayed. The selected score (here: weighted score) is highlighted in grey. PathoLive detects all 7 prevalent viruses as the top hits. Additionally, the visualization clearly emphasizes all spiked pathogens through the thickness of their clades, while other species are shown only in smaller clades and therefore ranked lower. The results do not change significantly in the later stages of the run.

While the total hits T_n can be useful to get a general impression of the abundance of sequences in the sample, the unambiguous bases U_n provides a first comparison to the background dataset. The weighted score W_n introduces an intensified metric of how often a sequence is found in a healthy individual, and thereby allows drawing stricter conclusions from the background data. Not only exactly overlapping mappings of fore-

and background are regarded, but also the overall abundance of a sequence within the background data is considered.

The values of the selected scoring scheme are reflected in the thickness of the branches, which draws the visual focus to higher-rated branches. Users can switch between the three scores via the respective buttons in the interactive visualization. In order to enable users to make early decisions regarding the handling of a sample as well as to further enhance the intuitive understanding of the results, the hazardousness of detected pathogens is color-coded based on a biosafety level (BSL) score list [56]. To improve BSL classification, minor changes were manually applied to improve matches to the organism names in the reference database. The BSL score gives information on the biological risk emanating from an organism. Therefore, it qualifies as a measure of hazardousness in this use case. The BSL score is color-coded in green (no information/BSL1), blue (BSL2), yellow (BSL3), or red (BSL4), and the maximum hazardousness level of a branch is propagated to the parent nodes. While the BSL level may be non-informative in some cases, it can still help in gaining an overview for non-expert users who are not familiar with certain pathogens. Phages are displayed in grey, as they cannot infect humans directly, but may imply information on the presence of bacteria.

Details about the sums of all three available scores of all underlying species are provided on mouse-over (Figure 3 in the results section). When expanding a branch to sequence level, additional plots of the foreground coverage calculated in step (iii) as well as the abundance of bases in the background datasets calculated in step (ii) are shown when hovering the mouse over the node (Figure 2). These plots provide a visualization of the significance of a hit. The hits of a species in the patient dataset are shown in green, while background hits are drawn in red on a coverage plot. This way, it is easy to evaluate if a sequence is commonly found in non-ill humans and therefore can be considered less relevant, or if a detected sequence is unique and could lead to more certain conclusions. The source code is available at https://gitlab.com/rki_bioinformatics/PathoLive, accessed on 23 August 2022.

2.2. Validation

We compared the results of PathoLive to two existing solutions, Clinical Pathoscope [26] and Bracken [57]. We selected Clinical Pathoscope for its very sophisticated read reassignment method, which promises a highly reliable rating of candidate hits. It also is perfectly tailored to this use case. Other promising pipelines such as SURPI [30] were not locally installable and had to be disregarded. Bracken, a method based on metagenomics classification with Kraken [22], was included in the benchmark as one of the fastest and best-known classification tools which makes it one of the primary go-to methods for many users. The experiment is based on a real sequencing run on an Illumina HiSeq 1500 in High Output Mode. We designed an in-house generated sample in order to have a solid ground truth. We ran all tools using 40 threads, starting each at the earliest possible time point when the data was available from the sequencer in the expected input format. For the non-real-time tools, the base calling was executed via Illumina's standard tool bcl2fastq and the runtime was regarded in the overall turnaround time. Clinical Pathoscope and Bracken were both run with default parameters, apart from the multithreading. We built the databases for PathoLive and Bracken using the viral part of the NCBI RefSeq [58]. For Clinical Pathoscope, we downloaded the associated database from http://www.bu.edu/jlab/wp-assets/databases.tar.gz (accessed on 23 August 2022) using the provided viral database as foreground and the human database as background. Details of the database construction are given in the Supplementary Methods (Section S3.1). Please note that, in contrast to all other results shown in this manuscript, the live analysis of the in-house sample was performed using read-mapping results of HiLive, the predecessor of HiLive2. However, we repeated the analysis using HiLive2 and obtained similar results with respect to accuracy (cf. Supplementary Figure S1 and Supplementary Table S1).

To validate PathoLive on real data, we applied it to a previously described diagnostic human serum sample from an outbreak of hemorrhagic fever virus in Sudan [59,60] and a dataset from an outbreak of Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in Wuhan, 2019 [61]. As the data were only available in FASTQ format, it was converted to BCL file format following the procedure described in the supplementary methods (Section S3.2). The total read length was 2 × 301 bp for the CCHFV dataset from Sudan and 2 × 151 bp for the SARS-CoV-2 dataset from Wuhan.

2.3. Sample Preparation

Viral metagenomics studies were performed with a human plasma mix of six different RNA and DNA viruses as well-defined surrogates for clinical liquid specimens. This 200 µL mix contained orthopoxvirus (Vaccinia virus VR-1536), flavivirus (yellow fever virus 17D vaccine), paramyxovirus (mumps virus vaccine), bunyavirus (rift valley fever virus MP12-vaccine), reovirus (T3/Bat/Germany/342/08), and adenovirus (human adenovirus 4) from cell culture supernatant at different concentrations. For the reovirus culture, we inoculated specific pathogen-free embryonated chicken eggs with reovirus as described by Kohl et al. [62]. All other viruses were passaged in HEp-2 or VerE6 cells. The cultured viruses were mixed with a human plasma sample. The CT values in the final sample ranged from 14.0 and 39.1. The sample also contained dependoparvovirus as proven via PCR. The sample was filtered through a 0.45 µM filter and nucleic acids were extracted using the QIAamp Ultrasense Kit (Qiagen, Venlo, tHE Netherlands) following the manufacturer's instructions. The extract was treated with Turbo DNA (Life Technologies, Darmstadt, Germany). cDNA and double-stranded cDNA (ds-cDNA) synthesis were performed as previously described (70). The ds-cDNA was purified with the RNeasy MinElute Cleanup Kit (Qiagen). The purification method takes ~6 h to complete. The library preparation was performed with the Nextera XT DNA Sample Preparation Kit following the manufacturer's instructions (Illumina, San Diego, CA, USA). NGS libraries were quantified using the KAPA Library Quantification Kits for Illumina sequencing (Kapa Biosystems, Wilmington, NC, USA). If the starting amount of 1 ng of nucleic acid was not reached. the entire sample volume was added to the library.

The diagnostic sample from Sudan was prepared according to [59,60], including inactivation of the human serum in Qiagen Buffer AVL, extraction with Qiagen QIAamp Viral RNA Mini Kit, and DNA digestion using the Thermo Fisher TURBO DNA-free Kit. A sequencing library was created using the Illumina Nextera XT DNA Library Preparation Kit. The sample was sequenced on an Illumina MiSeq.

The dataset from the outbreak of SARS-CoV-2 in Wuhan in 2019 was sequenced on an Illumina MiniSeq sequencing device and is publicly available at the NCBI Sequence Read Archive (SRA) under accession number SRR10971381 [61].

3. Results

3.1. Pathogen Detection in a Spiked Viral Mixture

The human plasma sample spiked with a viral mixture was sequenced on an Illumina HiSeq 1500 in high output mode on one lane. PathoLive was executed from the beginning of the sequencing run using 40 threads. Intermediary results were taken after 40, 60, 80, and 100 cycles or after 36, 55, 74, and 93 h, respectively. The time needed to produce results from the intermediary sequencing data was lower than 25 min for all output cycles. Raw reads usable for the testing of other tools were available only after 95 h as they had to be translated into the human-readable fastq-format first. As a ground truth, we selected all sequences associated with the species described as abundant above.

The area under the curve (AUC) of the receiver operating characteristic (ROC) was calculated using the 14 highest ranking species, as given by the tested tools. The top 14 of the identified species are considered because hits appearing after twice the number of true positives cannot be expected to be regarded by a user in this experiment. Furthermore, none of the tested tools found more true positives within the next 50 hits. The ROC plot

(Figure 4) denotes the true-positive rate and false-positive rate for each threshold $n \leq 14$, whereby a threshold n means that the best n hits are taken into account. This means that only the rank of the hits was considered while disregarding the actual score. For PathoLive, the ranks were determined by the weighted score W_n, for Clinical Pathoscope we used the "final guess" metric, and for Bracken, the species with the most estimated reads were ranked highest.
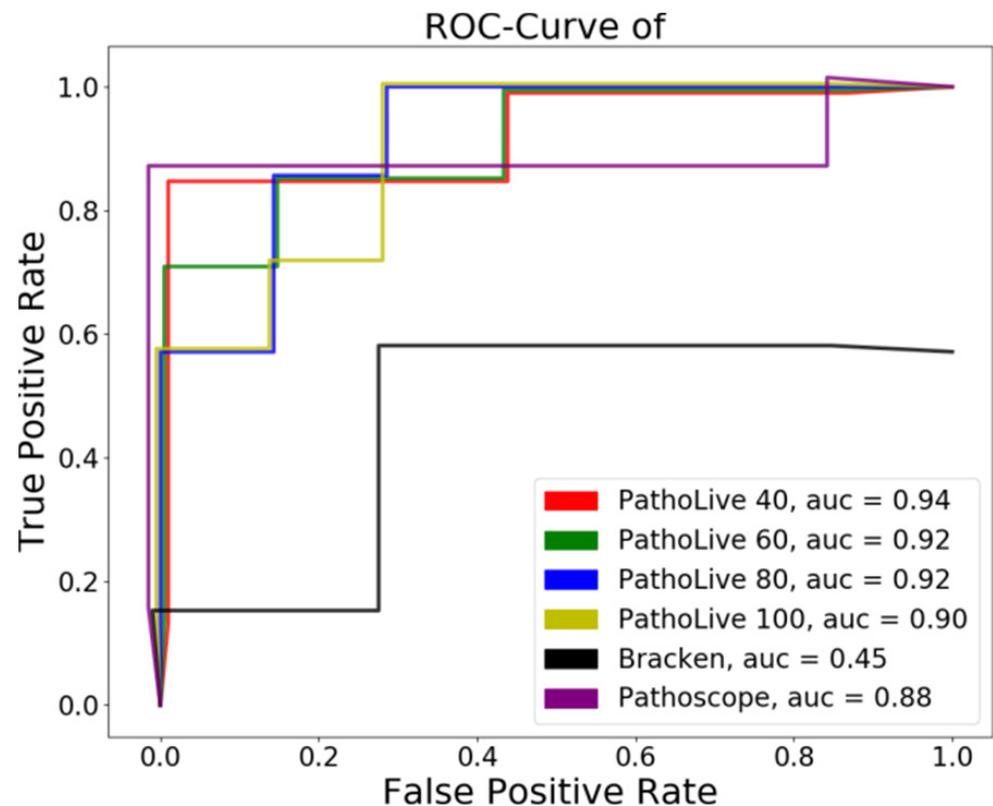

Figure 4. ROC-plot of benchmarked tools on a spiked dataset. Lines have slight offsets in x- and y-dimensions for reasons of distinguishability. We compared PathoLive to Clinical Pathoscope and Bracken on a human sample containing 7 viruses. PathoLive performs best regarding the ROC-AUC at all sampled times (cycle 40, 60, 80, and 100) when compared to the results of the other tools after sequencing the complete first read (cycle 100).

We were able to detect all abundant spiked species in the library after only 40 cycles of the sequencing run using PathoLive. While the overall number of false positive hits decreases with the sequencing time, the weighted score, and the number of unambiguous bases yield accurate results throughout all reports. Reported phages are included in these numbers, although they are optically grayed out in the visualization, as they cannot infect vertebrates directly. As an example report, a screenshot of the resulting interactive tree of results after 80 cycles is shown in Figure 3.

3.2. Identification of Crimean-Congo Hemorrhagic Fever Virus in a Real Sample from Sudan

A central issue in pathogen identification, especially for viruses, is the potentially low number of pathogenic reads in the sample. Therefore, we demonstrated the performance of PathoLive on real data that is known to contain a low number of reads of interest. We analyzed a human serum sample from Sudan that was confirmed via PCR to contain Crimean-Congo hemorrhagic fever virus (CCHFV) but only shows a small amount of related reads in the corresponding Illumina sequencing data (45 out of 1,178,054 reads were reported by Andrusch et al. in 2018 [59] to unambiguously belong to CCHFV). When running PathoLive with default parameters and having adapter trimming activated,

Bunyaviridae was the family with the highest weighted score over the complete sequencing procedure when not considering phages and the "unassigned family" branch. Thereby, the score of Bunyaviridae was consistently equal to the score of the underlying species CCHFV while other underlying species did not contribute to the overall score of the family. Figure 5 shows the development over time for all families that reach a score of 500 in at least one output cycle. It can be seen that the weighted score of CCHFV (represented by the family of Bunyaviridae) is in the top three of all identified families after only 30 sequencing cycles, which corresponds to 5% of the sequencing procedure. At this time point, only 16 reads were aligned to CCHFV. Thus, indications for the correct finding are already possible within a short time span and based on only a couple of available reads while the result is more and more emphasized with ongoing sequencing. The only other family reaching a score higher than 500 and not exclusively containing phages was Retroviridae, being mainly driven by the species HIV1. However, a more detailed view of the sequence level shows that all mappings to HIV1 cluster in a small region of approximately 1000 bp (Figure 6d) while the alignments to CCHFV distribute over the complete genome (Figure 6b,c). This strongly indicates that CCHFV is more likely to be a true positive. Figure 6 further shows the family-level visualization of the PathoLive tree structure (Figure 6a) and an example for Granulovirus of the Baculoviridae family that shows a high total number of mappings, but all of those being located in regions that are covered in the background database leading to a weighted score of 0 (Figure 6e). The overall results for this sample show the strength of PathoLive to pronounce interesting findings at first glance while still allowing for a more detailed perspective that is often important for interpretation.
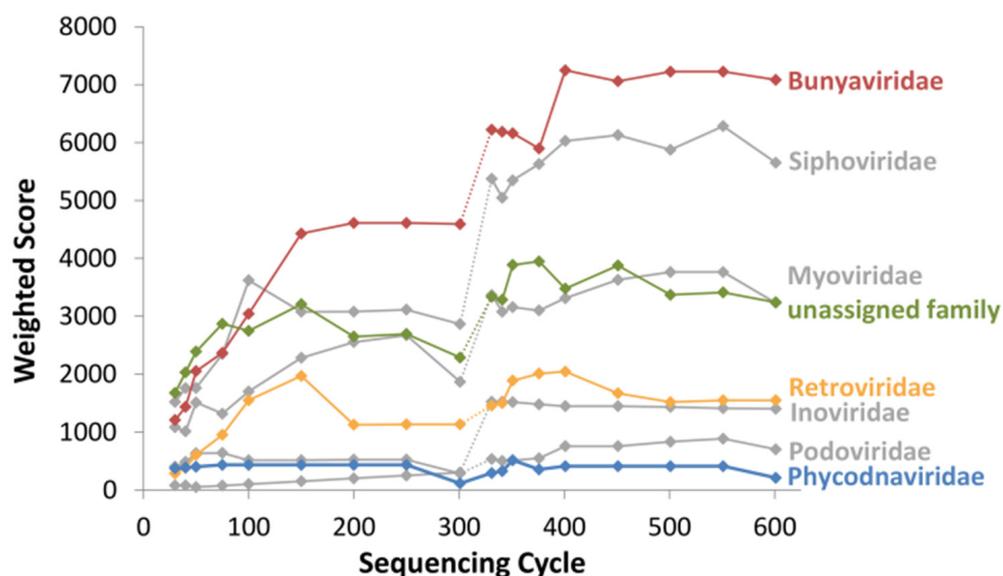

Figure 5. Development of the weighted score calculated by PathoLive on the real-world CCHFV sample from Sudan over the sequencing procedure for all families reaching a score higher than 500 in at least one output cycle. Colors of the plots correspond to the underlying biosafety level in the last cycle, i.e., green for BSL-1, blue for BSL-2, and yellow for BSL-3. Phages are displayed in gray color. The dotted section of each line indicates the shift from the first to the second read of the 2 × 301 bp data.
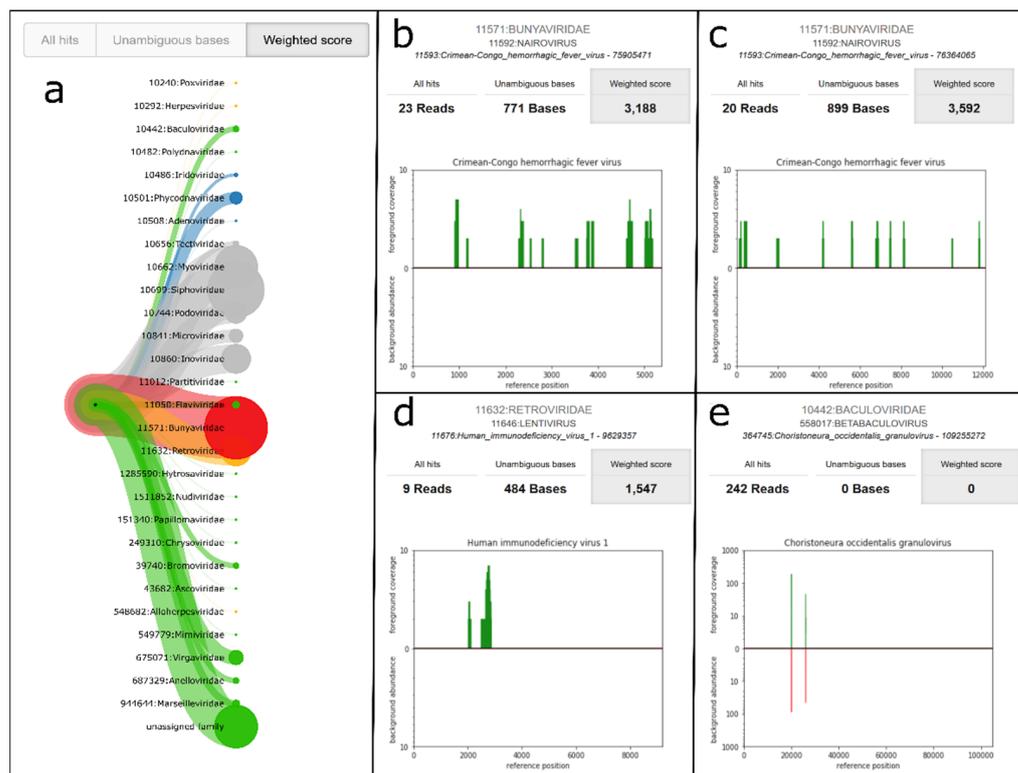
Figure 6. **Visualization of the final results of PathoLive for cycle 602.** (**a**) Tree structure on family level. (**b**,**c**) Tooltips for the sequence level of alignments for two CCHFV reference sequences of the Bunyaviridae family. (**d**) Tooltip for the sequence level of alignments for HIV1 of the Retroviridae family. (**e**) Tooltip for the sequence level of alignments for a Granulovirus reference of the Baculoviridae family.

3.3. Detection of a Coronavirus in a Real Sample from the 2019 SARS-CoV-2 Outbreak in Wuhan

For a dataset from the outbreak of SARS-CoV-2 in Wuhan, 2019, we could also identify a coronavirus as the most probable causative virus. This example clearly demonstrates the strength of our scoring approach.

When using the pure quantity of alignments for the visualization of results, the Coronaviridae family branch is not among the most prominently visualized hits. In contrast, when activating the weighted score, a clear indication was already available after only 30 sequencing cycles, corresponding to 10% of the complete sequencing run (Figure 7). The ranks and underlying scores of the visualization shown in Table 1 further support the strength of the weighted score approach. A more detailed analysis of the underlying tree of the Coronaviridae family shows that there are different coronavirus species with high scores, mainly dominated by a SARS-related coronavirus and a bat coronavirus. This further indicates that a clear assignment to one of the underlying species was not possible. These results are as expected, since the correct species, SARS-CoV-2, was not present in the reference database.

Another branch that is clearly highlighted by its red color belongs to the Poxviridae family. However, a more detailed look into the results shows that the BSL-4 classification originates from a single sub-branch where all mapping positions cluster to only two single peaks (not shown). This is a similar pattern to what we already showed for the occurrence of HIV1 in the previous section (cf. Figure 6d) and is therefore most probably not of biological interest.
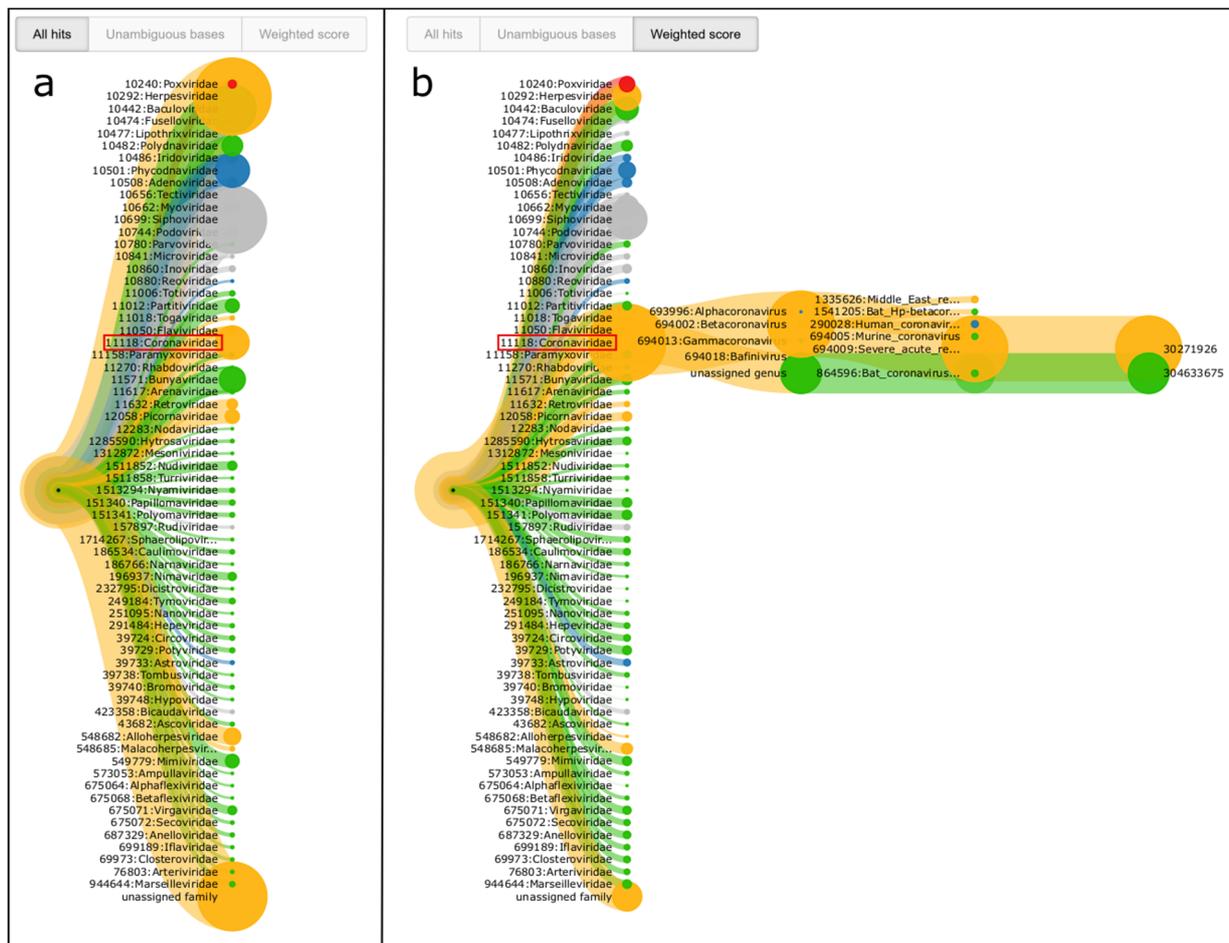
Figure 7. Visualization of the PathoLive results on a real-world dataset from the Wuhan 2019 SARS-CoV-2 outbreak after 30 sequencing cycles. (**a**) Branch thickness determined by the absolute number of aligned reads T_n. (**b**) By selecting the weighted score, the results correctly highlight the presence of a coronavirus by assigning the clearly highest score. When opening the details of the coronavirus branch, highest similarity can be determined with SARS-related coronavirus and a bat coronavirus. In both subfigures, the family of Coronaviridae is marked by a red box.

Table 1. Comparison of the ranking of the Coronaviridae family for total hits and weighted score. While the Coronaviridae family is not among the most abundant hits, it clearly shows the highest weighted score. The total ranking contains 67 families of which the six highest scores are shown.

	Total Hits			**Weighted Score**	
	Family	**Hits**		**Family**	**Score**
1	Herpesviridae	21,811	**1**	**Coronaviridae**	**46,274**
2	unassigned family	17,345	2	Siphoviridae	11,559
3	Siphoviridae	16,838	3	unassigned family	6148
4	Baculoviridae	7996	4	Herpesviridae	5342
5	Phycodnaviridae	4018	5	Myoviridae	3775
6	**Coronaviridae**	**3827**	6	Baculoviridae	3423
. . .			. . .		

4. Discussion

NGS has been shown to be the current state-of-the-art DNA sequencing technology for pathogen detection. Although it is still a niche technology in clinical settings due to

the described shortcomings, it makes an increasing impact on the diagnosis of infectious diseases. Although third-generation sequencing approaches are also becoming more and more influential, the discovery of lowly abundant pathogens is still problematic due to the relatively low number of reads. Additionally, the comparably low coverage and high error rates still hamper certain types of complex follow-up analyses such as the detection of antimicrobial resistances or the geographical origin of a pathogen. On the other hand, long-read sequencing technology shows immense potential for real-time diagnostics in the future, especially when considering the continuously decreasing error rates, shorter sample preparation times, arising higher throughput devices such as the PromethION, as well as valuable technology-specific features such as the read until functionality for first attempts that have been made to separate microbial reads from host DNA during the sequencing procedure [47,63]. All these aspects considered we assume long-read sequencing technology a valuable complement to NGS-based diagnostics in the future with distinct properties and therefore potentially different application areas.

The high turnaround time of NGS-based diagnostics is a major drawback compared to targeted molecular methods. Past efforts to speed up NGS-based diagnostics have been made but often come with significant disadvantages: Quick, Ashton, Calus, Chatt, Gossain, Hawker, Nair, Neal, Nye, Peters, De Pinna, Robinson, Struthers, Webber, Catto, Dallman, Hawkey and Loman [42] introduced a fast-sequencing protocol for Illumina sequencers that allows obtaining results after as little as 6 h. This speedup is accompanied by lower throughput and lower data quality, making it less suitable for whole genome shotgun sequencing approaches without a priori knowledge. Other approaches aiming at performing analyses of intermediate sequencing data require either a massive reduction of the amount of analyzed reads and/or targets [43] or the application of specialized hardware such as field-programmable gate array technology (FPGA) which is, for example, used for the DRAGEN system [44]. Such specialized hardware approaches come with additional costs, either for the purchase and infrastructure of local solutions or for the use of a cloud system. At the same time, such approaches provide a low level of flexibility in the analysis and are not algorithmically optimized for working with incomplete data. PathoLive does not require the use of specialized hardware and provides accurate diagnostics results in real time, illustrated with an easily understandable and interactive visualization. This strongly facilitates getting insights into a clinical sample before the sequencer has finished. Real-time output before the sequencing process of the first read has finished lacks information about multiplex indices, though. Therefore, early results of multiplexed sequencing runs can only be assigned to a specific sample after sequencing of the multiplex indices. For paired-end sequencing runs, this still means analyses are still possible far before the sequencer ends, and single-end sequencing runs can produce results at the very moment the indices have been sequenced. A possible solution for this problem is to sequence the indices before the first read, which can pose addressable challenges for cluster identification. As a working solution, many sequencing devices allow paired-end sequencing with different lengths for the first and second reads. It is thereby possible to sequence only a short fragment of the first read to get early access to the multiplex indices. Thus, this approach can be used to obtain de facto single-end reads (i.e., the full second read) while having the multiplex information available from the beginning of the read.

For pathogen identification, we changed the basis for the selection of clinically relevant hits from pure abundance or coverage-based measures towards a metric that takes information on the singularity of a detected pathogen into account. Still, we decided not to completely trust the algorithmic evaluation alone but provide all available information to the user in an intuitive interactive taxonomic tree. While we assume that this form of presentation allows users to come to the right conclusions very quickly, more sophisticated methods for abundance estimation especially on strain level exist. Implementing an additional abundance estimation approach comparable to the read reassignment of Clinical Pathoscope [26] or the abundance estimation of Bracken [57] could enable more accurate results, albeit this would not be applicable trivially to the overall conception of PathoLive.

The sensitivity and specificity of PathoLive vary with the time of a sequencing run. In the beginning, when only little sequence information is available, only a small number of nucleotides specify a candidate hit, leading to comparably high false positive rates. At the end of a sequencing run, the number of sequence mismatches in the longer alignments may lead to the erroneous exclusion of hits, especially when sequencing quality decreases. However, this behavior is implicitly considered by the HiLive2 algorithm which allows for an increasing number of mismatching nucleotides with the increasing length of the reads. Still, the results can vary over runtime with the optimal outcome being measured at intermediate cycles if the selected parameters are not well-suited for the specific sample or if the sequencing quality decreases stronger than usual.

While at this state PathoLive is focused on the detection of pathogenic viruses due to the design of the reference database, the concept of PathoLive is transferable to bacteria or other pathogens by using the database updater and rerunning the background definition.

Besides these challenges which are unique to PathoLive, similar problems as conventional approaches occur. First, the definition of meaningful reference databases is difficult. No reference database can ever be exhaustive since not all existing organisms have been sequenced yet. Besides that, there may be erroneous information in the reference databases due to sequencing artifacts, contaminations, or false taxonomic assignment. The definition of hazardousness was especially complicated, as to our knowledge no well-established solution for the automated assignment of this information exists. Therefore, the basis for our BSL-levelling approach might not be exhaustive, leading to underestimated danger levels of pathogens that are missing in the underlying BSL list. Furthermore, in-house contaminations, some of which are known to be carried over from run to run on the sequencer while others may come from the lab, could interfere with the result interpretation of a sequencing run. Especially since no indices are sequenced for the first results of PathoLive, comparably large numbers of carry-over contaminations might lead to false conclusions. Candidate contaminations should therefore be kept in mind when interpreting results.

We believe that PathoLive combines the qualities of different existing tools while adding several new, helpful features, such as live analyses and user-friendly visualizations. That way, we outperform the runtime of existing methods in an unprecedented manner and optimize the overall turnaround time from sample receipt to workable results. For retrospective analysis of previously sequenced samples, other workflows may offer a different scope or insights that are more detailed. However, to gain a quick overview of a sample, the tailored methods included in PathoLive offer an ideal solution.

Using in-house generated spiked human plasma samples, we were able to show the advantages of PathoLive not only concerning its unprecedented runtime but also the selection of relevant pathogens. We further show the high sensitivity of our approach by identifying CCHFV in a real sample from Sudan based on a few dozen reads. While being very fast and accurate, a limitation of PathoLive lies in the discovery of yet unknown pathogens. This is due to the limited sensitivity of alignment-based methods in general, which hampers the correct assignment of highly deviant sequences. However, the analysis of a dataset from the SARS-CoV-2 outbreak in 2019 clearly shows that the detection of novel species that are related to known pathogens is still possible. Concluding, PathoLive is a helpful tool for accurate and yet rapid detection of pathogens in clinical NGS datasets. The key advantages are the real-time availability of analysis results as well as the intuitive and interactive visualization with the down-prioritization of likely irrelevant candidates.

Author Contributions: Conceptualization, S.H.T., T.P.L., P.W.D., M.S.L., A.N. and B.Y.R.; data curation, S.H.T., T.P.L., J.K., P.W.D. and A.N.; formal analysis, S.H.T., T.P.L. and B.Y.R.; funding acquisition,

References

1. Bzhalava, D.; Johansson, H.; Ekstrom, J.; Faust, H.; Moller, B.; Eklund, C.; Nordin, P.; Stenquist, B.; Paoli, J.; Persson, B.; et al. Unbiased approach for virus detection in skin lesions. *PLoS ONE* **2013**, *8*, e65953. [CrossRef] [PubMed]
2. Greninger, A.L.; Zerr, D.M.; Qin, X.; Adler, A.L.; Sampoleo, R.; Kuypers, J.M.; Englund, J.A.; Jerome, K.R. Rapid Metagenomic Next-Generation Sequencing during an Investigation of Hospital-Acquired Human Parainfluenza Virus 3 Infections. *J. Clin. Microbiol.* **2017**, *55*, 177–182. [CrossRef] [PubMed]
3. Breitwieser, F.P.; Pardo, C.A.; Salzberg, S.L. Re-analysis of metagenomic sequences from acute flaccid myelitis patients reveals alternatives to enterovirus D68 infection. *F1000Research* **2015**, *4*, 180. [CrossRef] [PubMed]
4. Salzberg, S.L.; Breitwieser, F.P.; Kumar, A.; Hao, H.; Burger, P.; Rodriguez, F.J.; Lim, M.; Quinones-Hinojosa, A.; Gallia, G.L.; Tornheim, J.A.; et al. Next-generation sequencing in neuropathologic diagnosis of infections of the nervous system. *Neurol. Neuroimmunol. Neuroinflamm.* **2016**, *3*, e251. [CrossRef] [PubMed]
5. Cao, M.D.; Ganesamoorthy, D.; Elliott, A.G.; Zhang, H.; Cooper, M.A.; Coin, L.J. Streaming algorithms for identification of pathogens and antibiotic resistance potential from real-time MinION(TM) sequencing. *Gigascience* **2016**, *5*, 32. [CrossRef]
6. Roux, S.; Tournayre, J.; Mahul, A.; Debroas, D.; Enault, F. Metavir 2: New tools for viral metagenome comparison and assembled virome analysis. *BMC Bioinform.* **2014**, *15*, 76. [CrossRef]
7. Kostic, A.D.; Ojesina, A.I.; Pedamallu, C.S.; Jung, J.; Verhaak, R.G.; Getz, G.; Meyerson, M. PathSeq: Software to identify or discover microbes by deep sequencing of human tissue. *Nat. Biotechnol.* **2011**, *29*, 393–396. [CrossRef]
8. Skewes-Cox, P.; Sharpton, T.J.; Pollard, K.S.; DeRisi, J.L. Profile hidden Markov models for the detection of viruses within metagenomic sequence data. *PLoS ONE* **2014**, *9*, e105067. [CrossRef]
9. Wommack, K.E.; Bhavsar, J.; Polson, S.W.; Chen, J.; Dumas, M.; Srinivasiah, S.; Furman, M.; Jamindar, S.; Nasko, D.J. VIROME: A standard operating procedure for analysis of viral metagenome sequences. *Stand. Genom. Sci.* **2012**, *6*, 427–439. [CrossRef]
10. Dutilh, B.E.; Schmieder, R.; Nulton, J.; Felts, B.; Salamon, P.; Edwards, R.A.; Mokili, J.L. Reference-independent comparative metagenomics using cross-assembly: crAss. *Bioinformatics* **2012**, *28*, 3225–3231. [CrossRef]
11. Norling, M.; Karlsson-Lindsjo, O.E.; Gourle, H.; Bongcam-Rudloff, E.; Hayer, J. MetLab: An In Silico Experimental Design, Simulation and Analysis Tool for Viral Metagenomics Studies. *PLoS ONE* **2016**, *11*, e0160334. [CrossRef] [PubMed]
12. Huson, D.H.; Beier, S.; Flade, I.; Gorska, A.; El-Hadidi, M.; Mitra, S.; Ruscheweyh, H.J.; Tappu, R. MEGAN Community Edition—Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. *PLoS Comput. Biol.* **2016**, *12*, e1004957. [CrossRef] [PubMed]
13. Zhao, G.; Wu, G.; Lim, E.S.; Droit, L.; Krishnamurthy, S.; Barouch, D.H.; Virgin, H.W.; Wang, D. VirusSeeker, a computational pipeline for virus discovery and virome composition analysis. *Virology* **2017**, *503*, 21–30. [CrossRef]
14. Tausch, S.H.; Renard, B.Y.; Nitsche, A.; Dabrowski, P.W. RAMBO-K: Rapid and Sensitive Removal of Background Sequences from Next Generation Sequencing Data. *PLoS ONE* **2015**, *10*, e0137896. [CrossRef]
15. Piro, V.C.; Matschkowski, M.; Renard, B.Y. MetaMeta: Integrating metagenome analysis tools to improve taxonomic profiling. *Microbiome* **2017**, *5*, 101. [CrossRef] [PubMed]

16. Bray, N.L.; Pimentel, H.; Melsted, P.; Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **2016**, *34*, 525–527. [CrossRef]

17. Menzel, P.; Ng, K.L.; Krogh, A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat. Commun.* **2016**, *7*, 11257. [CrossRef]

18. Zheng, Y.; Gao, S.; Padmanabhan, C.; Li, R.; Galvez, M.; Gutierrez, D.; Fuentes, S.; Ling, K.S.; Kreuze, J.; Fei, Z. VirusDetect: An automated pipeline for efficient virus discovery using deep sequencing of small RNAs. *Virology* **2017**, *500*, 130–138. [CrossRef]

19. Dadi, T.H.; Renard, B.Y.; Wieler, L.H.; Semmler, T.; Reinert, K. SLIMM: Species level identification of microorganisms from metagenomes. *PeerJ* **2017**, *5*, e3138. [CrossRef] [PubMed]

20. Lee, A.Y.; Lee, C.S.; Van Gelder, R.N. Scalable metagenomics alignment research tool (SMART): A scalable, rapid, and complete search heuristic for the classification of metagenomic sequences from complex sequence populations. *BMC Bioinform.* **2016**, *17*, 292. [CrossRef]

21. Piro, V.C.; Lindner, M.S.; Renard, B.Y. DUDes: A top-down taxonomic profiler for metagenomics. *Bioinformatics* **2016**, *32*, 2272–2280. [CrossRef]

22. Wood, D.E.; Salzberg, S.L. Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* **2014**, *15*, R46. [CrossRef] [PubMed]

23. Truong, D.T.; Franzosa, E.A.; Tickle, T.L.; Scholz, M.; Weingart, G.; Pasolli, E.; Tett, A.; Huttenhower, C.; Segata, N. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* **2015**, *12*, 902–903. [CrossRef] [PubMed]

24. Scheuch, M.; Hoper, D.; Beer, M. RIEMS: A software pipeline for sensitive and comprehensive taxonomic classification of reads from metagenomics datasets. *BMC Bioinform.* **2015**, *16*, 69. [CrossRef]

25. Hong, C.; Manimaran, S.; Shen, Y.; Perez-Rogers, J.F.; Byrd, A.L.; Castro-Nallar, E.; Crandall, K.A.; Johnson, W.E. PathoScope 2.0: A complete computational framework for strain identification in environmental or clinical sequencing samples. *Microbiome* **2014**, *2*, 33. [CrossRef]

26. Byrd, A.L.; Perez-Rogers, J.F.; Manimaran, S.; Castro-Nallar, E.; Toma, I.; McCaffrey, T.; Siegel, M.; Benson, G.; Crandall, K.A.; Johnson, W.E. Clinical PathoScope: Rapid alignment and filtration for accurate pathogen identification in clinical samples using unassembled sequencing data. *BMC Bioinform.* **2014**, *15*, 262. [CrossRef] [PubMed]

27. Francis, O.E.; Bendall, M.; Manimaran, S.; Hong, C.; Clement, N.L.; Castro-Nallar, E.; Snell, Q.; Schaalje, G.B.; Clement, M.J.; Crandall, K.A.; et al. Pathoscope: Species identification and strain attribution with unassembled sequencing data. *Genome Res.* **2013**, *23*, 1721–1729. [CrossRef] [PubMed]

28. Flygare, S.; Simmon, K.; Miller, C.; Qiao, Y.; Kennedy, B.; Di Sera, T.; Graf, E.H.; Tardif, K.D.; Kapusta, A.; Rynearson, S.; et al. Taxonomer: An interactive metagenomics analysis portal for universal pathogen detection and host mRNA expression profiling. *Genome Biol.* **2016**, *17*, 111. [CrossRef]

29. Lindner, M.S.; Renard, B.Y. Metagenomic abundance estimation and diagnostic testing on species level. *Nucleic Acids Res.* **2013**, *41*, e10. [CrossRef]

30. Naccache, S.N.; Federman, S.; Veeraraghavan, N.; Zaharia, M.; Lee, D.; Samayoa, E.; Bouquet, J.; Greninger, A.L.; Luk, K.C.; Enge, B.; et al. A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. *Genome Res.* **2014**, *24*, 1180–1192. [CrossRef] [PubMed]

31. Piro, V.C.; Dadi, T.H.; Seiler, E.; Reinert, K.; Renard, B.Y. ganon: Precise metagenomics classification against large and up-to-date sets of reference sequences. *bioRxiv* **2019**, 406017. [CrossRef] [PubMed]

32. Wood, D.E.; Lu, J.; Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **2019**, *20*, 257. [CrossRef]

33. Breitwieser, F.P.; Lu, J.; Salzberg, S.L. A review of methods and databases for metagenomic classification and assembly. *Brief Bioinform* **2017**, *20*, 1125–1136. [CrossRef] [PubMed]

34. Dutilh, B.E.; Reyes, A.; Hall, R.J.; Whiteson, K.L. Editorial: Virus Discovery by Metagenomics: The (Im)possibilities. *Front. Microbiol.* **2017**, *8*, 1710. [CrossRef] [PubMed]

35. Frey, K.G.; Herrera-Galeano, J.E.; Redden, C.L.; Luu, T.V.; Servetas, S.L.; Mateczun, A.J.; Mokashi, V.P.; Bishop-Lilly, K.A. Comparison of three next-generation sequencing platforms for metagenomic sequencing and identification of pathogens in blood. *BMC Genom.* **2014**, *15*, 96. [CrossRef]

36. Lecuit, M.; Eloit, M. The diagnosis of infectious diseases by whole genome next generation sequencing: A new era is opening. *Front. Cell. Infect. Microbiol.* **2014**, *4*, 25. [CrossRef]

37. Lecuit, M.; Eloit, M. The potential of whole genome NGS for infectious disease diagnosis. *Expert. Rev. Mol. Diagn.* **2015**, *15*, 1517–1519. [CrossRef]

38. Mokili, J.L.; Rohwer, F.; Dutilh, B.E. Metagenomics and future perspectives in virus discovery. *Curr. Opin. Virol.* **2012**, *2*, 63–77. [CrossRef]

39. Roux, S.; Emerson, J.B.; Eloe-Fadrosh, E.A.; Sullivan, M.B. Benchmarking viromics: An in silico evaluation of metagenome-enabled estimates of viral community composition and diversity. *PeerJ* **2017**, *5*, e3817. [CrossRef]

40. Snyder, L.A.; Loman, N.; Pallen, M.J.; Penn, C.W. Next-generation sequencing–the promise and perils of charting the great microbial unknown. *Microb. Ecol.* **2009**, *57*, 1–3. [CrossRef]

41. Breitwieser, F.P.; Pertea, M.; Zimin, A.V.; Salzberg, S.L. Human contamination in bacterial genomes has created thousands of spurious proteins. *Genome Res.* **2019**, *29*, 954–960. [CrossRef] [PubMed]

42. Quick, J.; Ashton, P.; Calus, S.; Chatt, C.; Gossain, S.; Hawker, J.; Nair, S.; Neal, K.; Nye, K.; Peters, T.; et al. Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of Salmonella. *Genome Biol.* **2015**, *16*, 114. [CrossRef] [PubMed]

43. Stranneheim, H.; Engvall, M.; Naess, K.; Lesko, N.; Larsson, P.; Dahlberg, M.; Andeer, R.; Wredenberg, A.; Freyer, C.; Barbaro, M.; et al. Rapid pulsed whole genome sequencing for comprehensive acute diagnostics of inborn errors of metabolism. *BMC Genom.* **2014**, *15*, 1090. [CrossRef] [PubMed]

44. Miller, N.A.; Farrow, E.G.; Gibson, M.; Willig, L.K.; Twist, G.; Yoo, B.; Marrs, T.; Corder, S.; Krivohlavek, L.; Walter, A.; et al. A 26-hour system of highly sensitive whole genome sequencing for emergency management of genetic diseases. *Genome Med.* **2015**, *7*, 100. [CrossRef] [PubMed]

45. Tausch, S.H.; Strauch, B.; Andrusch, A.; Loka, T.P.; Lindner, M.S.; Nitsche, A.; Renard, B.Y. LiveKraken—-Real-time metagenomic classification of illumina data. *Bioinformatics* **2018**, *34*, 3750–3752. [CrossRef]

46. Greninger, A.L.; Naccache, S.N.; Federman, S.; Yu, G.; Mbala, P.; Bres, V.; Stryke, D.; Bouquet, J.; Somasekar, S.; Linnen, J.M.; et al. Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. *Genome Med.* **2015**, *7*, 99. [CrossRef]

47. Loose, M.; Malla, S.; Stout, M. Real-time selective sequencing using nanopore technology. *Nat. Methods* **2016**, *13*, 751–754. [CrossRef]

48. Stewart, R.D.; Watson, M. poRe GUIs for parallel and real-time processing of MinION sequence data. *Bioinformatics* **2017**, *33*, 2207–2208. [CrossRef]

49. Loka, T.P.; Tausch, S.H.; Renard, B.Y. Reliable variant calling during runtime of Illumina sequencing. *Sci. Rep.* **2019**, *9*, 16502. [CrossRef]

50. Brister, J.R.; Ako-Adjei, D.; Bao, Y.; Blinkova, O. NCBI viral genomes resource. *Nucleic Acids Res.* **2015**, *43*, D571–D577. [CrossRef]

51. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **2015**, *526*, 68–74. [CrossRef] [PubMed]

52. Bolger, A.M.; Lohse, M.; Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **2014**, *30*, 2114–2120. [CrossRef] [PubMed]

53. Langmead, B.; Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **2012**, *9*, 357–359. [CrossRef] [PubMed]

54. Lindner, M.S.; Renard, B.Y. Metagenomic profiling of known and unknown microbes with microbeGPS. *PLoS ONE* **2015**, *10*, e0117711. [CrossRef]

55. Bostock, M.; Ogievetsky, V.; Heer, J. D(3): Data-Driven Documents. *IEEE Trans. Vis. Comput. Graph.* **2011**, *17*, 2301–2309. [CrossRef]

56. Biosafety and Biotechnology Unit. Belgian Classifications for Micro-Organisms Based on Their Biological Risks—Definitions. 20087. Available online: https://my.absa.org/Riskgroups (accessed on 23 August 2022).

57. Lu, J.; Breitwieser, F.P.; Thielen, P.; Salzberg, S.L. Bracken: Estimating species abundance in metagenomics data. *PeerJ Computer Science* **2017**, *3*, e104. [CrossRef]

58. O'Leary, N.A.; Wright, M.W.; Brister, J.R.; Ciufo, S.; Haddad, D.; McVeigh, R.; Rajput, B.; Robbertse, B.; Smith-White, B.; Ako-Adjei, D.; et al. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **2016**, *44*, D733–D745. [CrossRef]

59. Andrusch, A.; Dabrowski, P.W.; Klenner, J.; Tausch, S.H.; Kohl, C.; Osman, A.A.; Renard, B.Y.; Nitsche, A. PAIPline: Pathogen identification in metagenomic and clinical next generation sequencing samples. *Bioinformatics* **2018**, *34*, i715–i721. [CrossRef]

60. Kohl, C.; Eldegail, M.; Mahmoud, I.; Schrick, L.; Radonic, A.; Emmerich, P.; Rieger, T.; Gunther, S.; Nitsche, A.; Osman, A.A. Crimean congo hemorrhagic fever, 2013 and 2014 Sudan. *Int. J. Infect. Dis.* **2016**, *53*, 9. [CrossRef]

61. Wu, F.; Zhao, S.; Yu, B.; Chen, Y.-M.; Wang, W.; Song, Z.-G.; Hu, Y.; Tao, Z.-W.; Tian, J.-H.; Pei, Y.-Y.; et al. A new coronavirus associated with human respiratory disease in China. *Nature* **2020**, *579*, 265–269. [CrossRef]

62. Kohl, C.; Brinkmann, A.; Dabrowski, P.W.; Radonic, A.; Nitsche, A.; Kurth, A. Protocol for metagenomic virus detection in clinical specimens. *Emerg. Infect. Dis.* **2015**, *21*, 48–57. [CrossRef] [PubMed]

63. Edwards, H.S.; Krishnakumar, R.; Sinha, A.; Bird, S.W.; Patel, K.D.; Bartsch, M.S. Real-Time Selective Sequencing with RUBRIC: Read Until with Basecall and Reference-Informed Criteria. *Sci. Rep.* **2019**, *9*, 11475. [CrossRef] [PubMed]