PathoLive - Real time pathogen identification from metagenomic Illumina datasets

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

[1] Centre for Biological Threats and Special Pathogens: Highly Pathogenic Viruses (ZBS 1), Robert Koch Institute, Berlin, Germany

[2] Bioinformatics Division (MF 1), Department for Methods Development and Research Infrastructure, Robert Koch Institute, Berlin, Germany

[3] Molecular Microbiology and Genome Analysis (Unit 46), Department Biological Safety, Federal Institute for Risk Assessment, Berlin, Germany.

Abstract

Over the past years, NGS has been applied in time critical applications such as pathogen diagnostics with promising results. Yet, long turnaround times have to be accepted to generate sufficient data, as the analysis can only be performed sequentially after the sequencing has finished. Additionally, the interpretation of results can be further complicated by various types of contaminations, clinically irrelevant sequences, and the sheer amount and complexity of the data.

We designed and implemented PathoLive, a real-time diagnostics pipeline which allows the detection of pathogens from clinical samples up to several days before the sequencing procedure is even finished and currently available tools may start to run. We adapted the core algorithm of HiLive, a real-time read mapper, and enhanced its accuracy for our use case. Furthermore, common contaminations, low-entropy areas, and sequences of widespread, non-pathogenic organisms are automatically marked beforehand using NGS datasets from healthy humans as a baseline. The results are visualized in an interactive taxonomic tree that provides an intuitive overview and detailed measures regarding the relevance of each identified potential pathogen.

We applied the pipeline on a human plasma sample that was spiked *in vitro* with *vaccinia virus, yellow fever virus, mumps virus, Rift Valley fever virus, adenovirus,* and *mammalian orthoreovirus*. The sample was then sequenced on an Illumina HiSeq. All spiked agents were detected after the completion of only 12% of the sequencing procedure and were ranked more accurately throughout the run than by any of the tested tools on the complete data. We also found a large number of other sequences and these were correctly marked as clinically irrelevant in the resulting visualization. This tagging allows the user to obtain the correct assessment of the situation at first glance.

PathoLive is available at https://gitlab.com/rki_bioinformatics/PathoLive.

1 Introduction

The ability to sequence large amounts of nucleic acids in an unbiased manner through NGS is particularly interesting for metagenomics studies. Metagenomic NGS has been proposed as a valuable technique for clinical application. Nucleic acids of pathogens can be detected in metagenomic clinical samples even in cases where routine procedures fail to identify the underlying causes of a patient's symptoms [1-4]. Most other pathogen detection methods such as polymerase chain reaction (PCR), cell culture, or amplicon sequencing, aim to detect predefined organisms. On the contrary, NGS facilitates the detection and even characterization of pathogens without *a priori* knowledge about candidate species. NGS, unlike any other method, generates sufficient data to detect even lowly abundant pathogens without targeted amplification of defined sequences. Thus, it allows for an unbiased diagnostic analysis.

There is a variety of tool able to address NGS-based pathogen related questions with different focuses: either aiming to discover yet unknown genomes [5-22] or to detect known species in a sample [23-40]. Among both groups, there are different underlying algorithms, the main distinction running between alignment-based [15-17, 19, 23, 25, 26, 28-31, 33, 35-37, 39, 40] and alignment-free methods [6, 9, 12, 21, 32, 38]. Many tools of course combine both approaches [5, 7, 8, 10, 11, 13, 14, 18, 20, 22, 27, 34]. While being faster in most cases, alignment-free methods are limited to the detection of sequences, whereas alignment-based methods potentially allow for a more detailed characterization of genomes.

Existing approaches based on unbiased full genome sequencing of metagenomic samples are facing various obstacles, especially concerning the ranking of the results according to their clinical relevance and the long overall turnaround time [41-48].

A central issue in NGS-based pathogen detection is that the clinically relevant data is very hard to identify. Not only is the host genome usually the dominating part in a metagenomic patient sample, but additionally there are nucleic acids of various clinically irrelevant species such as some *endogenous retroviruses* (ERV) or non-pathogenic bacteria which commonly colonize a person.

Even viruses may contain sequences of ERVs, as for example *gallid herpesvirus type 2* or *fowlpox virus*, potentially confusing the correct assignment of reads [49]. For these reasons, the number of reads hinting towards a relevant pathogen can be very limited and even be as low as a handful of individual reads. To compensate for the overwhelming amount of background sequences without introducing unwanted biases and thus risking a loss of signal, large numbers of reads are necessary. Still, there is no guarantee to get a sufficiently high coverage for the detection of a targeted pathogen genome.

To put it more generally, it is a widespread misconception to rely only on quantitative measures when ranking the importance of candidate hits. While the amount of nucleic acids of a pathogen in a sample may correlate with the phase or intensity of an infection, it may not be sufficient to select the most abundant species as the causative pathogen. On the contrary, not the amount but the uncommonness of a species in a given sample may give decisive 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. To

date, there are several pipelines tackling this problem in different ways. Many pathogen detection pipelines propose to define a reference database of host and contaminating sequences [10, 12, 27, 36, 38, 40]. While facilitating cleaner results, it may lead to a premature rejection of relevant sequences. The definition of precise contamination databases proves rather difficult and has not yet been adequately solved. Thus, deletion of relevant hits and misinterpretation of irrelevant hits still remains a common problem.

Generally, handling high numbers of detected species with a low number of reads each makes it very difficult to get a clear definition of relevant and irrelevant hits. A presentation of all detected hits without any weighting would be hard to interpret, wasting precious time at the end of the workflow. Yet, deleting any results to gain a better overview comes at great risk of overlooking the true cause of an infection. Not only background and contamination removal introduces the risk of losing information that might be relevant in the following diagnostic process. Intensity filters, as implemented e.g. in SLIMM [28], disregard sequences with too small genome coverages. As the author states, this step eliminates many genomes. This problem even intensifies for marker-gene based methods such as MetaPhlAn2 [33], 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 actually relevant candidates.

Moreover, sequencing and analyzing the necessary amount of data is very time consuming. An Illumina HiSeq run in High Output Mode, potentially necessary to detect lowly abundant viruses, takes up to 11 days. Thus, in urgent cases or acute outbreak situations, standard workflows take too long to generate results in time to take the necessary measures. There is a plethora of infectious diseases which can be lethal, especially if not treated timely. For example, ebola patients who die from the disease die after 9.8 ± 0.7 days after the first symptoms occur on average [50]. To obtain actionable results within an appropriate time frame to help these patients and to prevent further spreading of the disease, it is crucial to reduce the time span of the entire workflow from sample receipt to complete diagnosis.

Efforts to speed up NGS based diagnostics have been made but come with significant disadvantages: Quick et al. introduced a fast sequencing protocol for Illumina sequencers that allows obtaining results after as little as 6 hours [51]. 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.

There are several promising approaches of pathogen detection using the MinION handheld device for in field studies. While allowing impressive 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 [51-55].

Higher read numbers are indispensable for reliable pathogen detection. Therefore, the development of efficient methods to generate, analyze and understand large metagenomics datasets in an accurate and quick manner is crucial if NGS is to become 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 at the same time 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 common contaminations, background data and irrelevant species all at once. Tackling the problem of slow overall turnaround times, we applied and enhanced our in-house developed real-time read mapper HiLive that enables analyzing sequencing data while an Illumina sequencer is still running [56].

2 Methods

2.1 Implementation

In order to generate a quick, easy and robust pathogen diagnostics workflow, we implemented PathoLive. 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)** adapt HiLive, a real-time read mapper, to yield robust results even before the sequencer finishes, **(iv)** identify the hazardousness 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 sections.

i. Prepare reference databases to be more efficient in runtime

In order to save computational effort during the post-processing of the live-mapped reads, 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 [57], 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 . The obtained data are then merged. 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 the source organism of the sequence.
Afterwards, user definable subdatabases 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.

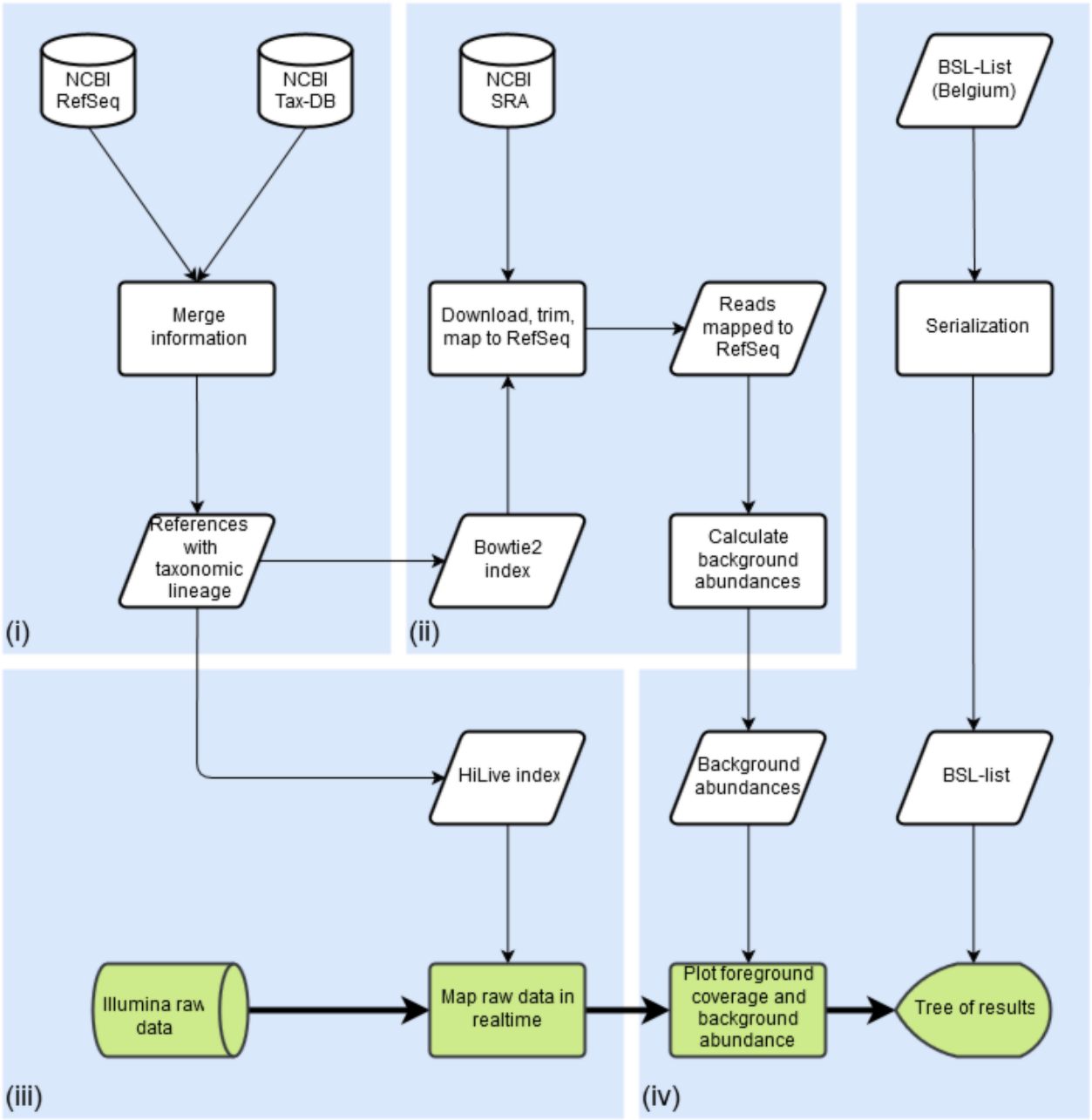

Figure 1: Workflow of PathoLive including four main modules: (i) Reference information from NCBI RefSeq is automatically downloaded and tagged with taxonomic information; (ii) NGS datasets from the 1000 Genomes Project are downloaded, trimmed and searched for sequences from the pathogen database from step (i), marking abundant stretches as clinically irrelevant; (iii) Reads from the clinical sample are mapped to the pathogen database obtained from (i) in real-time, producing intermediate alignment files in the bam-format at predefined time points; (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. Mark clinically irrelevant hits

A main obstacle in NGS based diagnostics is the large amount of background noise contained in the data. In this context, this refers to various sources of contamination including artificial sequences, 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 those sequences defined as irrelevant from the set of references carries the risk of losing ambiguous but 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. 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 [58] (s. Supplementary material) were downloaded, assuming that a large majority of the participants in the 1000 Genomes Project was not acutely ill with an infectious disease. These reads are quality trimmed using Trimmomatic [59] and mapped to the selected pathogen reference database using Bowtie2 [60]. 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 hit at least by one dataset are stored in the serialized pickle file format. Stretches of DNA found in this data are marked as clinically irrelevant and visualized as such in further steps of the workflow. The coverage maps of the background abundances are thereto plotted in red against the coverage maps of the reads from the patient dataset in green on the same reference (s. 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.

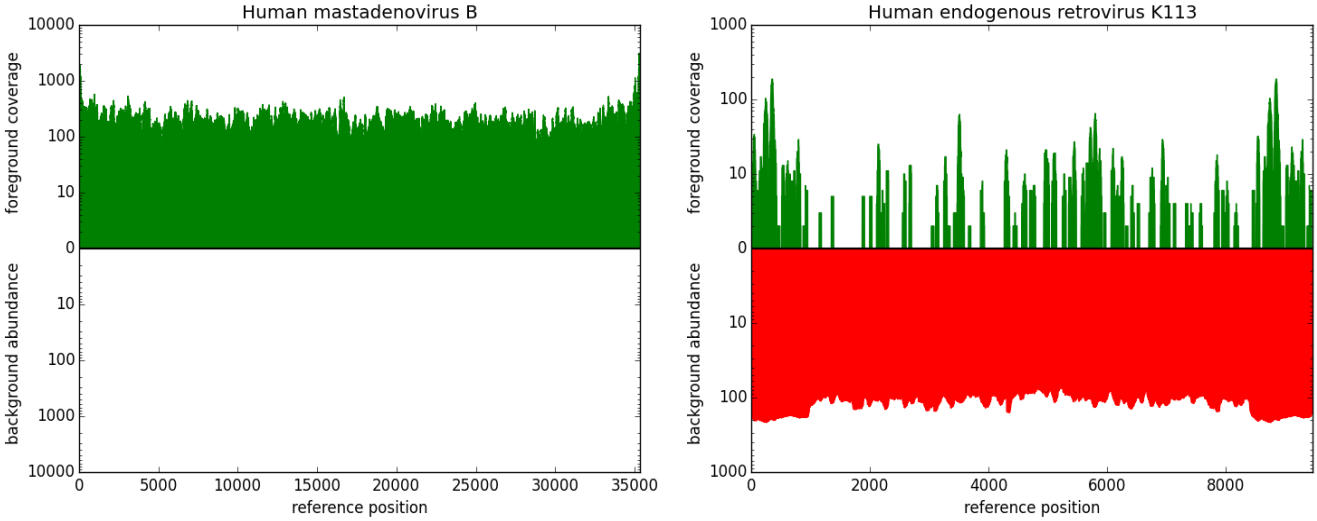

Figure 2: Two examples of fore- and background coverage plots. The upper, green bars show the coverage of a given genome in the foreground dataset, namely the reads sequenced from the patient sample. The lower, red part indicates in how many datasets from the 1000 Genomes Project a sequence is abundant. 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 abundance in non-ill humans.

iii. Adapt HiLive, Enhance to get results before sequencing finished

Due to the runtime requirements already mentioned, we aimed at breaking the sequential paradigm of wet and dry lab applications by parallelizing data generation and analysis. We used the real-time read mapper HiLive which yields results by the end of the sequencing run. To alleviate the high computational requirements to align all reads in parallel as they are sequenced, HiLive makes use of a highly efficient *k-mer* seed-and-extend approach. Therefore, errorless *k-mers* are looked up in a hash index. Each entry in the index contains matching positions for a *k-mer* in the database of reference genomes. Based on these *k-mer* positions, the q-gram lemma is applied to decide whether a certain *k-mer* position will be used to create, extend or discard an alignment candidate, referred to as seed [56]. Thereby, the user can decide how many errors to tolerate in an alignment. The algorithm results in a set of alignments for each read, including information about the matching genome and position but potentially missing detailed alignment information for regions with an accumulation of errors [56].

For the purpose of pathogen detection, we extended the current version (HiLive v0.3) by several features, resulting in a new version (HiLive v1.1). Instead of only obtaining results by the end of the sequencing run, HiLive now also contains the option to provide intermediate results at any point of a sequencing run with negligible delay. For the first time, this functionality allows not only to obtain mapping results at the same time the sequencing finishes but already during sequencing. The output of the mapping results was parallelized to handle even huge amounts of seeds that usually arise during intermediate steps. Additionally, we modified existing and created new output filters to reduce the number of random hits in the resulting alignment files. A separated executable can be used to create the output with different filter settings without re-executing the complete alignment algorithm. To further improve sensitivity, especially for the

mapping results in early sequencing cycles, we adapted the core algorithm to support arbitrary gapped *k-mers*. This means that single or consecutive mismatches are tolerated within a single *k-mer*. As shown by Kucherov et al. [61], this concept results in significantly higher accuracy especially after few cycles of a sequencing run, even though the q-gram lemma does not hold for gapped *k-mers*. For our study, we used SpEEd to select an optimal *k-mer* gap pattern for seeds of weight 15 and an expected similarity of 0.95 on 40 basepairs, resulting in the pattern 11111100111101 [62].
PathoLive is implemented in a modular manner. Instead of the real-time read mapping using HiLive, any other read mapper providing sequence alignment/map (sam) or binary sequence alignment/map (bam) files can be used for the mapping step of the workflow.

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 [40, 63] or interactive taxonomy explorers [16, 38]. 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 visualization library D3 [64]. For an example of the visualization, see Figure 4. While providing all available information on demand, the structure of a taxonomic tree allows an intuitive overview at first glance. Detailed measures are available on genus, family, species and sequence level. We provide three scores for each node of the tree:

(a) Total Hits: the total number of hits to all underlying sequences in this branch,

$$Total\ Hits = \#\ Reads\ mapped\ to\ clade$$

(b) Unambiguous Bases: the total number of bases covered in the patient dataset but not in any background dataset

$$Unambigious\ Bases = \#Bases\ covered\ by\ foreground\ but\ not\ by\ background\ data$$

(c) Weighted Score: the ratio of Unambiguous Bases to the number of bases covered by background reads

$$Weighted\ Score = \frac{Unambiguous\ Bases}{max(\#\ Bases\ covered\ by\ background\ data,1)} \times \log(Total\ Hits)\ .$$

The weighted score introduces an intensified metric of how often a sequence is found in non-ill persons, therefore allowing drawing stricter conclusions from the background data. Not only exactly overlapping mappings of fore- and background are regarded, but the overall abundance of a sequence within the background data is considered.

The values of these scores are reflected in the thickness of the branches, which draws the visual focus to higher rated branches. By default, the visualization uses the weighted score, but users can switch between all three scores.

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 [65]. 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. 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 4). When expanding a branch down 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 thus provide an intuitive visualization of the significance of a hit. The hits of a species in the patient dataset are shown in green while very common genomes or parts of their sequences are drawn in red on a correlating 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 very unique and could therefore lead to more certain conclusions.

2.2 Validation

We compared the results of PathoLive to two existing solutions, Clinical Pathoscope [36] and Bracken [66]. 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 [40] or Taxonomer [38] were not locally installable and had to be disregarded. Bracken 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 cores, 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 BaseCalling 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. The reference databases for PathoLive was built from the viral part of the NCBI RefSeq [67] downloaded on 2016-07-06. For Clinical Pathoscope we downloaded the associated database from http://www.bu.edu/jlab/wp-assets/databases.tar.gz on 2017-12-09 and used the provided viral database as foreground and the human database as background. The results of Bracken were generated based on the viral

part of the NCBI RefSeq [67] downloaded on 2017-12-18. The Bracken database was generated with default parameters and an expected read length of 100 bp.

Sample preparation

Viral RNA metagenomics studies were performed with a human plasma mix of eight different RNA and DNA viruses as well-defined surrogate for clinical liquid specimen. The informed consent of the patient has been obtained. This 200µL 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. The sample also contains *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) following the manufacturers' 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 [68]. The ds-cDNA was purified with the RNeasy MinElute Cleanup Kit (Qiagen). The purification method takes ~6h to complete.
The Library preparation was performed with the Nextera XT DNA Sample Preparation Kit following the manufacturers' instructions (Illumina). NGS libraries were quantified using the KAPA Library Quantification Kits for Illumina sequencing (Kapa Biosystems). If the starting amount of 1 ng of nucleic acid was not reached the entire sample volume was added to the library.

3 Results

The human plasma sample spiked with a viral mixture was subjected to sequencing 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 hours, respectively. Raw reads usable for the testing of other tools were available only after 95 hours as they had to be translated into the human readable fastq-format first. As a ground truth, we selected all sequences associated to the species described as abundant above. Turnaround time, runtime and results are shown in

Table 1. The area under the curve (auc) of the receiver operating characteristic (ROC) was calculated using the 16 highest ranking species, as given by the tested tools. The scores of all sequences attributed to a species were summed up. The top 16 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. For PathoLive, the weighted score is used, for Clinical Pathoscope we used the “final guess” metric and for Bracken, the species with most estimated reads were ranked highest. The corresponding ROC-plot is shown in Figure 6.

Table 1 Results of PathoLive, Clinical Pathoscope and Bracken on an Illumina HiSeq High Output run of a human plasma sample spiked with different viruses. Input data denotes the number of cycles the sequencer finished before results were generated. The turnaround time specifies the complete runtime of the sequencing from start of the sequencer to result presentation, whereas tool runtime is the time the tools take to generate results after all necessary input data has been provided. ROC-auc denotes the area under the ROC-curve as a combined measure of sensitivity and specificity. Best values are printed bold. PathoLive performs best according to all measures throughout the complete run.

	PathoLive				Pathoscope	Bracken
Input data [cycles]	40	60	80	100	100	100
Turnaround time [h]	**36**	55	74	93	95	95
Tool runtime [m]	22	25	18	**4**	25	13
ROC-auc	**0.94**	0.92	0.92	0.90	0.88	0.45

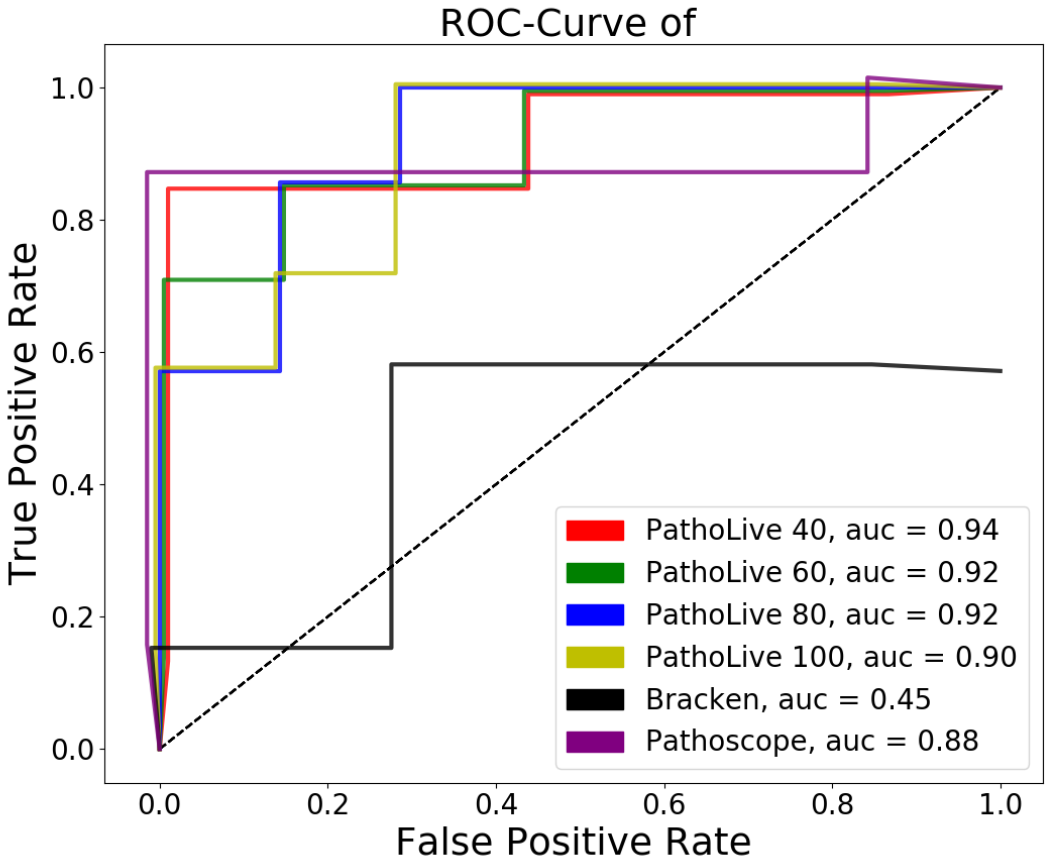

Figure 3 ROC-plot of benchmarked tools on a spiked dataset. Lines have slight offsets in x- and y-dimensions for reasons of distinguishability. We compared PathoLive to Clinical Pathoscope and Bracken on a real human sample containing 7 viruses. PathoLive performs best regarding the ROC-auc at all sampled times (cycle 40, 60, 80 and 100) when compared to the results of the other tools after the sequencing run completed read 1 (cycle 100).

We were able to detect all abundant spiked species in the library after only 40 cycles of the sequencing run. 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 4.

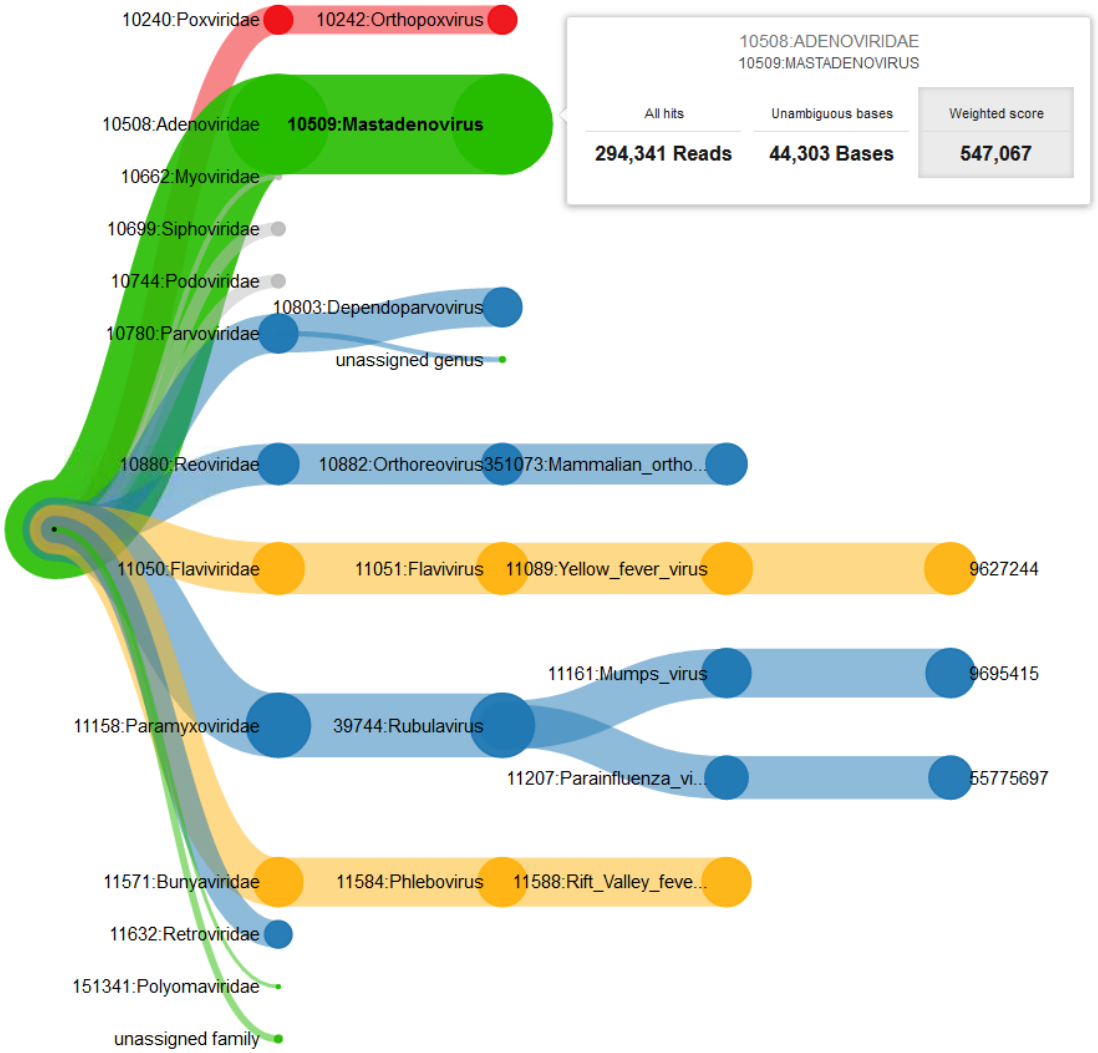

Figure 4 Example of the interactive taxonomic tree of results. It shows the visualized results of the described plasma sample at cycle 80 based on the weighted score. Thickness of the branches denotes the sum of scores of underlying sequences. The color codes for the maximum of the underlying BLS-levels (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. 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.

4 Discussion

NGS has been shown to be state of the art for pathogen detection, reaching out into clinical usage as well. Although Third Generation Sequencing approaches are also becoming more and more influential, the sequencing depth necessary for open-view diagnostics is only achievable via NGS. This does of course come at cost of higher overall throughput times. PathoLive is, to our knowledge, the first NGS-based diagnostics tool using a real-time approach, facilitating to gain 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, multiplexed sequencing runs can only be assessed 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 solution for this problem would be to sequence the indices before the first read, which attracts some problems for the sequencer regarding cluster identification, but is currently worked on. The algorithmic functionality for this is already available.

We furthermore changed the basis for the selection of clinically relevant pathogens away 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 the abundance estimation especially on strain level exist. Implementing an additional abundance estimation approach comparable to the read reassignment of Clinical Pathoscope [36] or the abundance estimation of Bracken [66] could enable more accurate results, albeit this would not be applicable trivially to the overall conception of PathoLive.

The sensitivity and specificity of PathoLive varies with the time of a sequencing run. In the beginning, when only little sequence information is available, every matching *k-mer* must be regarded as a candidate hit, leading to comparably high false positive rates. At the end of a sequencing run on the contrary, the number of sequence mismatches in the longer alignments may lead to the erroneous exclusion of hits. To cope with that, we recommend running PathoLive allowing high numbers of errors to ensure sensitive results at the end of a run and to report only reads with a low error-per-base ratio to exclude random hits at the beginning. This may however lead to the effect observed in our validation experiment, where the results vary over the runtime with the optimal outcome being measured at cycle 80.

Besides these challenges which are unique to PathoLive, we do of course struggle with the same problems as comparable tools. Firstly, 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 the hazardousness was especially complicated, as to our knowledge no 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 certain pathogens.

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 lab contaminations should therefore be thoroughly kept in mind when interpreting results.

Using in-house generated spiked human plasma samples, we were able to show the superiority of PathoLive not only concerning its unprecedented runtime but also the selection of relevant pathogens. 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. As this would imply tedious manual curation, it is not the core task of this tool.

We hope to provide a helpful tool for accurate and yet rapid detection of pathogens in clinical NGS datasets, overcoming many limitations of existing approaches.

5 References

1. Bzhalava, D., et al., *Unbiased approach for virus detection in skin lesions.* PLoS One, 2013. **8**(6): p. e65953.
2. Greninger, A.L., et al., *Rapid Metagenomic Next-Generation Sequencing during an Investigation of Hospital-Acquired Human Parainfluenza Virus 3 Infections.* J Clin Microbiol, 2017. **55**(1): p. 177-182.
3. Breitwieser, F.P., C.A. Pardo, and S.L. Salzberg, *Re-analysis of metagenomic sequences from acute flaccid myelitis patients reveals alternatives to enterovirus D68 infection.* F1000Res, 2015. **4**: p. 180.
4. Salzberg, S.L., et al., *Next-generation sequencing in neuropathologic diagnosis of infections of the nervous system.* Neurol Neuroimmunol Neuroinflamm, 2016. **3**(4): p. e251.
5. Li, Y., et al., *VIP: an integrated pipeline for metagenomics of virus identification and discovery.* Sci Rep, 2016. **6**: p. 23774.
6. Roux, S., et al., *Metavir 2: new tools for viral metagenome comparison and assembled virome analysis.* BMC Bioinformatics, 2014. **15**: p. 76.
7. Roux, S., et al., *Metavir: a web server dedicated to virome analysis.* Bioinformatics, 2011. **27**(21): p. 3074-5.
8. Kostic, A.D., et al., *PathSeq: software to identify or discover microbes by deep sequencing of human tissue.* Nat Biotechnol, 2011. **29**(5): p. 393-6.
9. Skewes-Cox, P., et al., *Profile hidden Markov models for the detection of viruses within metagenomic sequence data.* PLoS One, 2014. **9**(8): p. e105067.
10. Wommack, K.E., et al., *VIROME: a standard operating procedure for analysis of viral metagenome sequences.* Stand Genomic Sci, 2012. **6**(3): p. 427-39.
11. Zhao, G., et al., *Identification of novel viruses using VirusHunter--an automated data analysis pipeline.* PLoS One, 2013. **8**(10): p. e78470.
12. Dutilh, B.E., et al., *Reference-independent comparative metagenomics using cross-assembly: crAss.* Bioinformatics, 2012. **28**(24): p. 3225-31.

13. Norling, M., et al., *MetLab: An In Silico Experimental Design, Simulation and Analysis Tool for Viral Metagenomics Studies.* PLoS One, 2016. **11**(8): p. e0160334.
14. Alves, J.M., et al., *GenSeed-HMM: A Tool for Progressive Assembly Using Profile HMMs as Seeds and its Application in Alpavirinae Viral Discovery from Metagenomic Data.* Front Microbiol, 2016. **7**: p. 269.
15. Huson, D.H., et al., *MEGAN analysis of metagenomic data.* Genome Res, 2007. **17**(3): p. 377-86.
16. Huson, D.H., et al., *MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data.* PLoS Comput Biol, 2016. **12**(6): p. e1004957.
17. Huson, D.H. and S. Mitra, *Introduction to the analysis of environmental sequences: metagenomics with MEGAN.* Methods Mol Biol, 2012. **856**: p. 415-29.
18. Deng, X., et al., *An ensemble strategy that significantly improves de novo assembly of microbial genomes from metagenomic next-generation sequencing data.* Nucleic Acids Res, 2015. **43**(7): p. e46.
19. Huson, D.H. and N. Weber, *Microbial community analysis using MEGAN.* Methods Enzymol, 2013. **531**: p. 465-85.
20. Zhao, G., et al., *VirusSeeker, a computational pipeline for virus discovery and virome composition analysis.* Virology, 2017. **503**: p. 21-30.
21. Tausch, S.H., et al., *RAMBO-K: Rapid and Sensitive Removal of Background Sequences from Next Generation Sequencing Data.* PLoS One, 2015. **10**(9): p. e0137896.
22. Piro, V.C., M. Matschkowski, and B.Y. Renard, *MetaMeta: integrating metagenome analysis tools to improve taxonomic profiling.* Microbiome, 2017. **5**(1): p. 101.
23. Bray, N.L., et al., *Near-optimal probabilistic RNA-seq quantification.* Nat Biotechnol, 2016. **34**(5): p. 525-7.
24. Menzel, P., K.L. Ng, and A. Krogh, *Fast and sensitive taxonomic classification for metagenomics with Kaiju.* Nat Commun, 2016. **7**: p. 11257.
25. Naeem, R., M. Rashid, and A. Pain, *READSCAN: a fast and scalable pathogen discovery program with accurate genome relative abundance estimation.* Bioinformatics, 2013. **29**(3): p. 391-2.
26. Freitas, T.A., et al., *Accurate read-based metagenome characterization using a hierarchical suite of unique signatures.* Nucleic Acids Res, 2015. **43**(10): p. e69.
27. Zheng, Y., et al., *VirusDetect: An automated pipeline for efficient virus discovery using deep sequencing of small RNAs.* Virology, 2017. **500**: p. 130-138.
28. Dadi, T.H., et al., *SLIMM: species level identification of microorganisms from metagenomes.* PeerJ, 2017. **5**: p. e3138.
29. Lee, A.Y., C.S. Lee, and R.N. Van Gelder, *Scalable metagenomics alignment research tool (SMART): a scalable, rapid, and complete search heuristic for the classification of metagenomic sequences from complex sequence populations.* BMC Bioinformatics, 2016. **17**: p. 292.
30. Fosso, B., et al., *MetaShot: an accurate workflow for taxon classification of host-associated microbiome from shotgun metagenomic data.* Bioinformatics, 2017. **33**(11): p. 1730-1732.
31. Piro, V.C., M.S. Lindner, and B.Y. Renard, *DUDes: a top-down taxonomic profiler for metagenomics.* Bioinformatics, 2016. **32**(15): p. 2272-80.
32. Wood, D.E. and S.L. Salzberg, *Kraken: ultrafast metagenomic sequence classification using exact alignments.* Genome Biol, 2014. **15**(3): p. R46.
33. Truong, D.T., et al., *MetaPhlAn2 for enhanced metagenomic taxonomic profiling.* Nat Methods, 2015. **12**(10): p. 902-3.
34. Scheuch, M., D. Hoper, and M. Beer, *RIEMS: a software pipeline for sensitive and comprehensive taxonomic classification of reads from metagenomics datasets.* BMC Bioinformatics, 2015. **16**: p. 69.
35. Hong, C., et al., *PathoScope 2.0: a complete computational framework for strain identification in environmental or clinical sequencing samples.* Microbiome, 2014. **2**: p. 33.

36. Byrd, A.L., et al., *Clinical PathoScope: rapid alignment and filtration for accurate pathogen identification in clinical samples using unassembled sequencing data.* BMC Bioinformatics, 2014. **15**: p. 262.
37. Francis, O.E., et al., *Pathoscope: species identification and strain attribution with unassembled sequencing data.* Genome Res, 2013. **23**(10): p. 1721-9.
38. Flygare, S., et al., *Taxonomer: an interactive metagenomics analysis portal for universal pathogen detection and host mRNA expression profiling.* Genome Biol, 2016. **17**(1): p. 111.
39. Lindner, M.S. and B.Y. Renard, *Metagenomic abundance estimation and diagnostic testing on species level.* Nucleic Acids Res, 2013. **41**(1): p. e10.
40. Naccache, S.N., et al., *A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples.* Genome Res, 2014. **24**(7): p. 1180-92.
41. Breitwieser, F.P., J. Lu, and S.L. Salzberg, *A review of methods and databases for metagenomic classification and assembly.* Brief Bioinform, 2017.
42. Dutilh, B.E., et al., *Editorial: Virus Discovery by Metagenomics: The (Im)possibilities.* Front Microbiol, 2017. **8**: p. 1710.
43. Frey, K.G., et al., *Comparison of three next-generation sequencing platforms for metagenomic sequencing and identification of pathogens in blood.* BMC Genomics, 2014. **15**: p. 96.
44. Lecuit, M. and M. Eloit, *The potential of whole genome NGS for infectious disease diagnosis.* Expert Rev Mol Diagn, 2015. **15**(12): p. 1517-9.
45. Lecuit, M. and M. Eloit, *The diagnosis of infectious diseases by whole genome next generation sequencing: a new era is opening.* Front Cell Infect Microbiol, 2014. **4**: p. 25.
46. Mokili, J.L., F. Rohwer, and B.E. Dutilh, *Metagenomics and future perspectives in virus discovery.* Curr Opin Virol, 2012. **2**(1): p. 63-77.
47. Roux, S., et al., *Benchmarking viromics: an in silico evaluation of metagenome-enabled estimates of viral community composition and diversity.* PeerJ, 2017. **5**: p. e3817.
48. Snyder, L.A., et al., *Next-generation sequencing--the promise and perils of charting the great microbial unknown.* Microb Ecol, 2009. **57**(1): p. 1-3.
49. Niewiadomska, A.M. and R.J. Gifford, *The extraordinary evolutionary history of the reticuloendotheliosis viruses.* PLoS Biol, 2013. **11**(8): p. e1001642.
50. Schieffelin, J.S., et al., *Clinical illness and outcomes in patients with Ebola in Sierra Leone.* N Engl J Med, 2014. **371**(22): p. 2092-100.
51. Quick, J., et al., *Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of Salmonella.* Genome Biol, 2015. **16**: p. 114.
52. Cao, M.D., et al., *Streaming algorithms for identification of pathogens and antibiotic resistance potential from real-time MinION(TM) sequencing.* Gigascience, 2016. **5**(1): p. 32.
53. Greninger, A.L., et al., *Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis.* Genome Med, 2015. **7**: p. 99.
54. Loose, M., S. Malla, and M. Stout, *Real-time selective sequencing using nanopore technology.* Nat Methods, 2016. **13**(9): p. 751-4.
55. Stewart, R.D. and M. Watson, *poRe GUIs for parallel and real-time processing of MinION sequence data.* Bioinformatics, 2017. **33**(14): p. 2207-2208.
56. Lindner, M.S., et al., *HiLive: real-time mapping of illumina reads while sequencing.* Bioinformatics, 2017. **33**(6): p. 917-319.
57. Brister, J.R., et al., *NCBI viral genomes resource.* Nucleic Acids Res, 2015. **43**(Database issue): p. D571-7.
58. Genomes Project, C., et al., *A global reference for human genetic variation.* Nature, 2015. **526**(7571): p. 68-74.
59. Bolger, A.M., M. Lohse, and B. Usadel, *Trimmomatic: a flexible trimmer for Illumina sequence data.* Bioinformatics, 2014. **30**(15): p. 2114-20.

60. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nat Methods, 2012. **9**(4): p. 357-9.
61. Kucherov, G., L. Noe, and M. Roytberg, *Multiseed lossless filtration.* IEEE/ACM Trans Comput Biol Bioinform, 2005. **2**(1): p. 51-61.
62. Ilie, L., S. Ilie, and A.M. Bigvand, *SpEED: fast computation of sensitive spaced seeds.* Bioinformatics, 2011. **27**(17): p. 2433-4.
63. Lindner, M.S. and B.Y. Renard, *Metagenomic profiling of known and unknown microbes with microbeGPS.* PLoS One, 2015. **10**(2): p. e0117711.
64. Bostock, M., V. Ogievetsky, and J. Heer, *D(3): Data-Driven Documents.* IEEE Trans Vis Comput Graph, 2011. **17**(12): p. 2301-9.
65. Unit., B.a.B., *Belgian classifications for micro-organisms based on their biological risks - Definitions*. 2008: https://my.absa.org/Riskgroups.
66. Lu, J., F.P. Breitwieser, and S.L. Salzberg, *Bracken: estimating species abundance in metagenomics data.* PeerJ Computer Science. **3**(e104).
67. O'Leary, N.A., et al., *Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation.* Nucleic Acids Res, 2016. **44**(D1): p. D733-45.
68. Klenner, J., et al., *Comparing Viral Metagenomic Extraction Methods.* Curr Issues Mol Biol, 2017. **24**: p. 59-70.

Funding

The authors gratefully acknowledge financial support from the German Federal Ministry of Health [2515NIK043].

Competing interests

The authors declare no competing interests.