Virology Journal

RESEARCH

Open Access

Anellovirus abundance as an indicator for viral metagenomic classifier utility in plasma samples

Gabriel Montenegro de Campos[1], Luan Gaspar Clemente[2], Alex Ranieri Jerônimo Lima[3], Eleonora Cella[4], Vagner Fonseca[5,6], João Paulo Bianchi Ximenez[7], Milton Yutaka Nishiyama Jr[8], Enéas de Carvalho[9], Sandra Coccuzzo Sampaio[3], Marta Giovanetti[10,11], Maria Carolina Elias[3] and Svetoslav Nanev Slavov[3*]

Abstract

Background Viral metagenomics has expanded significantly in recent years due to advancements in next-generation sequencing, establishing it as the leading method for identifying emerging viruses. A crucial step in metagenomics is taxonomic classification, where sequence data is assigned to specific taxa, thereby enabling the characterization of species composition within a sample. Various taxonomic classifiers have been developed in recent years, each employing distinct classification approaches that produce varying results and abundance profiles, even when analyzing the same sample.

Methods In this study, we propose using the identification of Torque Teno Viruses (TTVs), from the *Anelloviridae* family, as indicators to evaluate the performance of four short-read-based metagenomic classifiers: Kraken2, Kaiju, CLARK and DIAMOND, when evaluating human plasma samples.

Results Our results show that each classifier assigns TTV species at different abundance levels, potentially influencing the interpretation of diversity within samples. Specifically, nucleotide-based classifiers tend to detect a broader range of TTV species, indicating higher sensitivity, while amino acid-based classifiers like DIAMOND and CLARK display lower abundance indices. Interestingly, despite employing different algorithms and data types (protein-based vs. nucleotide-based), Kaiju and Kraken2 performed similarly.

Conclusion Our study underscores the critical impact of classifier selection on diversity indices in metagenomic analyses. Kaiju effectively assigned a wide variety of TTV species, demonstrating it did not require a high volume of reads to capture diversity. Nucleotide-based classifiers like CLARK and Kraken2 showed superior sensitivity, which is valuable for detecting emerging or rare viruses. At the same time, protein-based approaches such as DIAMOND and Kaiju proved robust for identifying known species with low variability.

Keywords Taxonomic classifiers, Torque teno viruses, TTV, Metagenomics, Abundance

*Correspondence:
Svetoslav Nanev Slavov
svetoslav.slavov@fundacaobutantan.org.br
Full list of author information is available at the end of the article

Introduction

A primary function of metagenomic next-generation sequencing (mNGS) is the characterization of microbial abundance including protozoa, bacteria and viruses, across diverse sample types, including clinical specimens [1]. In the context of etiological diagnosis of infectious diseases, mNGS is particularly valuable for its ability to detect unsuspected or emerging viruses, owing to the unbiased and comprehensive nature of this process [2–4]. Historically, viral metagenomics has been concentrated on short-read classification due to the higher depths of the sequencing and the low viral loads usually present in clinical samples [5, 6]. Due to the short fragment size generated by mNGS, a primary challenge lies in achieving robust and reliable taxonomic classification of the obtained reads [6–9]. Several computational approaches have been developed to surpass these challenges, improve the accuracy of viral metagenomic data classification and better estimate the taxon abundance [7, 10, 11].

Many programs and computational pipelines are available for metavirome analysis and can be divided based on short and long reads. Such an abundance of taxonomic classifiers exhibits significant variations in key metrics including accuracy, speed, and computational resource requirements [7, 12]. Additional obstacles may include the high variability of taxonomic data, the limited comprehensiveness and static nature of databases used in classification, as well as substantial computational memory usage [7]. Most classifiers use similarity-based methods (homology and alignment) or composition-based approaches (oligonucleotide frequencies or k-mer matching) [12], with notable differences observed between them [7, 10–13].

A suitable virus model for evaluating the classification potential of taxonomic classifiers is represented by the anelloviruses, specifically the torque teno viruses (TTVs). These viruses are universally present, highly prevalent, and consistently rank among the most abundant findings in human metagenomic surveys [14–16]. TTV are not known to be associated with clinical diseases and are considered a commensal component of the human virome. To date, TTV is represented by 29 species. However, this number is relatively low given the frequent discovery of novel species through metagenomics [17]. A limited set of TTV reference genomes may affect classification accuracy, potentially leading to misassignments and the underrepresentation of certain lineages. However, numerous TTV species absent from reference databases may still be present in clinical samples. Therefore, a primary focus is to determine whether the taxonomic classifiers, based on their algorithms, would assign a greater or lesser number of known TTV species.

We hypothesized that the universal presence of TTV in the metavirome, along with its high genetic variability, could correlate to the effectiveness of taxonomic classifiers in accurately identifying viruses. Anelloviruses are suitable for classification purposes compared to other commensal viruses, such as human pegivirus-1 (HPgV-1) and bacteriophages. HPgV-1 exhibits a lower prevalence that varies depending on the tested group, being higher in high-risk patients and lower in the general population [18, 19]. Bacteriophages display a higher abundance in stool samples compared to plasma [20], but they belong to diverse families and subfamilies with varying genomic organization and lengths. This diversity renders them less suitable for use as classification controls, particularly in the context of plasma samples. In contrast, anelloviruses are nearly universally present across the population, making them more appropriate and specific for classification purposes.

Therefore, in this study, we evaluated four different short-read classifiers: nucleotide-based: Kraken2 [21, 22] and CLARK [23], protein-based: Kaiju [24] and DIAMOND [25]. These classifiers were chosen based on their widespread use in metagenomic studies and their reliance on distinct algorithms and databases, providing a comprehensive comparison of their performance in handling highly variable viral genomes. Using a universally present virus, like TTV, we assessed how each classifier handles viral diversity and accurately estimated TTV abundance in metagenomic samples.

Materials and methods

Clinical samples and raw sequence data

For this study, we used three raw sequence datasets from our previously published studies [26, 27]. The metagenomic analysis was conducted on plasma samples pooled together to reduce sequencing costs. Reverse transcription, amplification, and library preparation were performed as previously described [28]. The raw reads were generated by Illumina NextSeq 1000/2000 sequencer using the P3 flow cell. The sample pools were named AL1, PR2, and PR3. Pool AL1 contained plasma samples from acutely infected patients with inconclusive amplification profiles for dengue (DENV), zika (ZIKV) and chikungunya (CHIKV) viruses [26]. Pools PR2 and PR3 consisted of pooled plasma samples obtained from patients with prostate cancer [27]. Plasma samples were specifically chosen to standardize the study within a particular sample type, ensure the presence of TTVs, and take advantage of our prior experience working with them.

Processing of the raw sequencing data

The raw sequencing data was processed by a pipeline composed of the following main steps: quality control

and pre-processing, mapping the filtered sequences and taxonomic classification (Fig. 1). This in-house established pipeline has been deposited online at https://github.com/gmcampos27/MetaviromePipeline.

In brief, raw sequencing data were assessed for their quality using FastQC v.0.11.8 [29]. Trimming, filtering, and adaptor removal were performed using fastp v.0.20.0 [30]. Host depletion was performed in silico through human read mapping using the Burrows-Wheeler Aligner (BWA) v.0.7.17-r1188 [31]. We applied the BWA-MEM algorithm with the *Homo sapiens* genome (NCBI GRCh38.p14) as the reference. Unmapped Filtered reads (non-human reads) were subsequently classified taxonomically by the four classifiers included in this study.

Reference database

The choice of database significantly impacts the taxonomic profile, as using a comprehensive database compared to an optimized subset can lead to varying assignments, even with identical settings. To focus solely on the classifier's influence, we utilized the same database across all classifiers: NCBI RefSeq Viruses. Since this is a pre-existing database, installation guidelines are available on the official pages of each classifier (Supplement Table 1).

Metagenomic classifiers

We selected the metagenomic classifiers based on the following criteria: (i) ability to generate customized databases, (ii) open-source accessibility, (iii) applicability for pathogen detection (principally viruses), and (iv) the type of sequence database utilized. We selected two classifiers that incorporate nucleotide databases, Kraken v.2.1.3 and CLARK v.1.3.0.0, and two employing peptide databases, Kaiju v.1.8.2 and DIAMOND v.2.0.14.152, in addition, to access metagenomic profile, we use it alongside with MEGAN v.6.22.2 [32]. The classifiers used are not exclusive to plasma samples, as they are designed for the analysis of any sample type, including environmental or clinical samples, irrespective of the virus abundance present. More information regarding the characteristics of the classifiers can be found in Supplement Table 1. Profiling commands are provided in Supplementary Material 1.

Diversity analysis

To assess the performance of the classifiers, we compared the TTV relative abundance per classifier and calculated alpha diversity indexes (Shannon, Simpson, and Richness). We examined the classifiers' similarity using the obtained abundance data by applying hierarchical clustering with Euclidean distance. We used two metrics: one representing the total number of reads detected and another indicating virus presence or absence. These data were presented in heatmaps, in which the read counts were normalized using a z-score for better visualization.

Finally, we generated rarefaction curves for each classifier to evaluate the viral richness in general and more

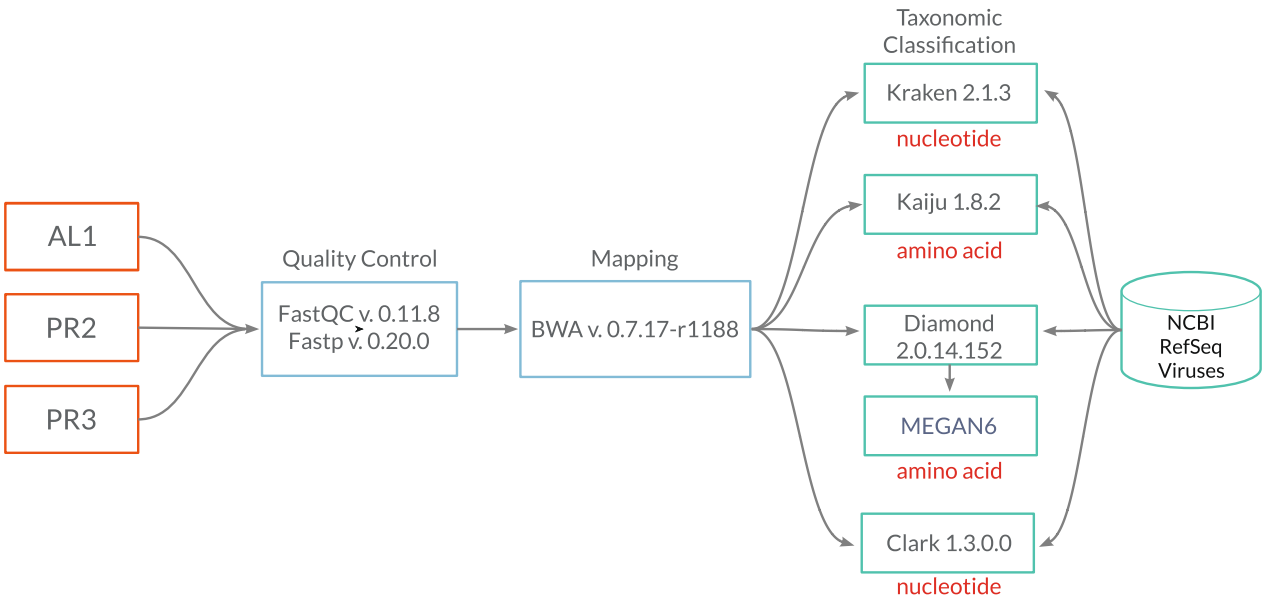

Fig. 1 Pipeline used in this study. Each raw sample progresses through three main stages: Quality Control using FastQC and fastp, Mapping using BWA, and Taxonomic Classification. For the classifier DIAMOND, an additional program, MEGAN6 is necessary to generate abundance profiles, as DIAMOND alone does not produce them. This workflow ensures that all sequences are processed equally

specifically for the TTV species. All analyses were performed using R v.4.3.1 [33] with the packages vegan v.2.6-6 [34] and ggplot2 v.3.5.1 [35], colors were provided using wesanderson color palette v.0.3.7 [36].

Statistical analysis

To evaluate differences in sequence reads assigned to TTVs across all classifiers, we performed the two-tailed Mann–Whitney U test ($\alpha = 0.05$), a non-parametric statistical test. To account for multiple comparisons, p-values were corrected using the Benjamini-Hochberg (BH) method. Read counts were previously normalized using log2 (counts per mil reads, CPM) to improve visualization and maintain statistical consistency. CPM normalization was chosen due to the fact it accounts for sequencing depth, facilitating comparisons across samples.

Results

Number of viral reads assigned by each classifier

The generated pools had a mean of 92,181,048 total reads. The quantitative characteristics of the mNGS after trimming and retrieving the human and unmapped reads are shown in Table 1. Information regarding the total number of reads was retrieved from the original articles [26, 27].

A comparison of the three sample pools (AL1, PR2, and PR3) revealed distinct patterns in the assignment of viral and TTV reads, highlighting variations in the performance of each classifier.

CLARK showed consistent behavior across all three samples, assigning 1,400,546 viral reads in each sample and identifying 18,801 TTV reads, regardless of the sample analyzed. However, while the total number of TTV reads was the same, the distribution of TTVs varied between the samples. On the other hand, Kraken2 assigned a very low number of viral reads in PR2, with 301 reads, but 261 reads were attributed to TTVs. In PR3 and AL1, the number of viral reads increased to 1,134 and 6,112, respectively. However, this increase did not correlate with TTV reads, as Kraken2 detected 882 TTV reads in PR3 and only 150 in AL1.

Kaiju demonstrated variable performance across the analyzed samples. In sample AL1, it detected 11,413 viral reads and 288 TTV reads. However, in PR2 and PR3, these numbers increased to 50,575 and 55,232 viral reads, with 280 and 457 TTV reads, respectively. These differences may reflect variations in the metagenomic composition of the samples. Similarly, Diamond also exhibited notable variation across the three samples. In AL1, Diamond detected 1,003,610 viral reads, including 19,708 TTV reads. In PR2 and PR3, the number of viral reads decreased to 32,998 and 139,113, respectively, while TTV reads dropped to 130 and 560. Given the universal presence of TTVs, we further evaluated their abundance across the included samples using different classifiers to better understand their classification capabilities.

Pool AL1: undetermined arboviruses

The analysis of the abundance of Pool AL1 revealed the presence of a high number of reads related to arboviruses (DENV type 2 and CHIKV), as stated in Souza et al., 2022. Moreover, a high diversity of commensal viruses was observed, across different subspecies of TTV. CLARK provided the highest TTV abundance in the sample AL1 (Fig. 2A). This classifier detected the highest number of unique *Alphatorquevirus* species (Fig. 2A), along with the highest TTV richness value (25 - Fig. 2G). The low Simpson diversity in this pool for CLARK, compared with a relatively high Shannon diversity index, suggests that the distribution of species is not even and

Table 1 Quantitative representation of the obtained reads from the pools

Pool name	Total read number	Reads after trimming	Human reads	Unmapped reads
AL1 [23]	101,593,212	46,754,909	8,527,762	279,814
PR1 [24]	88,341,428	86,320,012	58,487,186	27,755,645
PR2 [24]	86,608,503	85,432,732	58,458,464	26,841,137

(See figure on next page.)

Fig. 2 Comparison of TTV abundance and diversity across different taxonomic classifiers for AL1. **A** Relative abundance of TTVs identified by each classifier, with values ranging from 0 to 1.0 on the y-axis. on the y-axis. **B** Z-score normalized heatmap displaying read counts per classifier, with Kaiju forming an outer cluster. **C** Qualitative heatmap of the detected species, clustering similar-based classifiers; the x-axis represents classifiers, while the y-axis lists viral species, including unclassified *Alphatorquevirus*. **D** Natural log-transformed boxplot of read counts, with Clark yielding the highest number of reads. **E**–**G** Shannon, Simpson, and Richness diversity indices, respectively. **H** Rarefaction curves, with solid lines representing all viruses and dotted lines indicating only TTV species. **I** Same as (**H**), but with the x-axis limited to 30,000 reads for a clearer view of the initial rarefaction pattern

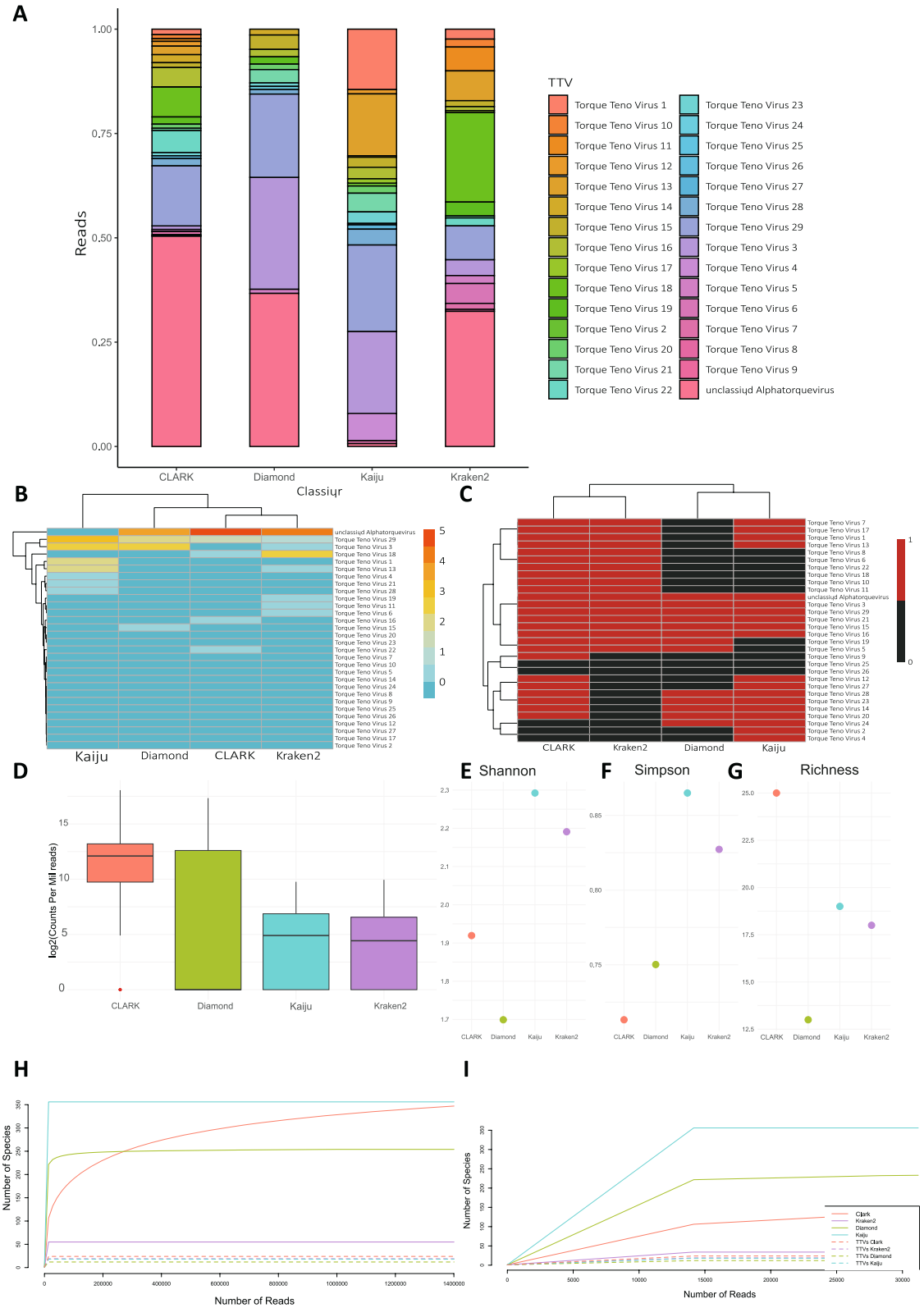

Fig. 2 (See legend on previous page.)

there are some species that are dominant in terms of frequency. When considering reads assigned to the TTV subspecies altogether, the classifiers based on nucleotide sequences—Kraken2 and CLARK—clustered closely, followed by DIAMOND and Kaiju, which formed a distantly located cluster (Fig. 2B). This distinct classification pattern is evident in Fig. 2B, where Kaiju exhibits strong signal intensities for multiple TTVs, particularly for TTV 29. This elevated presence potentially skews the overall uniformity of the sample, as it also suggests a higher Simpson value obtained by Kaiju (Fig. 2E). Additionally, Kaiju identified the lowest number of unclassified alphatorqueviruses, further supporting its distinct clustering behavior. When classifiers were grouped only by the absence and presence of TTV, classifiers of the same type were grouped (Fig. 2C).

Regarding the number of reads, CLARK attributed the highest number as belonging to TTVs, followed by DIAMOND, while Kraken2 and Kaiju produced comparably lower read counts (Fig. 2D). This distribution suggests that despite classifying a lower number of reads, Kaiju effectively assigned a wide variety of TTV species. CLARK generated a significantly higher number of TTV reads, indicating its capacity to detect a broader range of TTV sequences than other classifiers. Statistical analysis using the Mann–Whitney U Test confirmed the significance of this difference, with p-values below the stipulated alpha (0.05) for all comparative analyses involving CLARK (Table 2). This observation suggests that CLARK might potentially be more sensitive in detecting the total of the TTV species.

Table 2 Mann–Whitney Test results for AL1

Classifier 1	Classifier 2	*p*-value-adjusted*
Kraken 2	Kaiju	7.146874e−01
Kraken 2	DIAMOND	7.146874e−01
Kraken 2	CLARK	2.413528e−05
Kaiju	DIAMOND	7.146874e−01
Kaiju	CLARK	2.413528e−05
DIAMOND	CLARK	8.133686e−02

* *p*-value adjusted using BH method

The construction rarefaction curve illustrated the observed differences in the TTV classification potential of the used classifiers (Fig. 2H, I). CLARK demonstrated the slowest rarefaction trend, indicating that it required a higher read count to reach a plateau, reflecting its ability to identify a higher number of TTV species. Consistent with previous observations, Kraken2 and Kaiju were similar in terms of generating TTV abundance; however, Kaiju slightly outperformed Kraken2 by identifying a broader range of species overall. DIAMOND, while showing the fewest numbers of TTV species detections (protein-based), demonstrated relatively high species counts when considering the total viral diversity.

Pool PR2: plasma samples from prostate cancer patients

The main finding in the sample PR2 highlights a substantial presence of sequences belonging to Hepatitis C virus (HCV), alongside commensal viruses, such as TTVs [27]. Regarding the TTVs, in our study, DIAMOND identified only TTV 16, with other sequences being classified as unclassified alphatorqueviruses, which resulted in a lower value across all diversity indices (Fig. 3A).

While CLARK and Kraken2 demonstrated the highest TTV species richness [20], identifying a broad range of TTVs (Fig. 3G), their performance differed in other alpha diversity indices (Fig. 3E, F). CLARK exhibited lower Simpson and Shannon index values, suggesting that although both classifiers detected a similar number of species, Kraken2 assigned reads more evenly across taxa, whereas CLARK concentrated them in fewer species. This distinction influenced their clustering patterns, with CLARK grouping closely with DIAMOND in Fig. 3B due to the prevalence of uncharacterized alphatorqueviruses. In contrast, Kraken2 and Kaiju formed a separate cluster, likely reflecting similarities in read distribution patterns. As shown in Fig. 3C, G, Kraken2 and CLARK initially clustered together due to their shared richness levels, but broader clustering patterns later emerged, incorporating Kaiju and eventually DIAMOND.

Regarding the number of TTV reads assigned by each classifier, all performed similarly except for DIAMOND (Fig. 3D, H, I – dotted lines). Consequently, the Mann–Whitney U Test revealed significant differences only

(See figure on next page.)

Fig. 3 Comparison of TTV abundance and diversity across different taxonomic classifiers for PR2. **A** Relative abundance of TTVs identified by each classifier, with relative abundance values from 0 to 1.0 on the y-axis. **B** Z-score normalized heatmap of read counts, showing two main clusters: one formed by Kaiju and Kraken2, and another by CLARK and DIAMOND. **C** Qualitative heatmap of the detected species, DIAMOND was the outermost cluster because it did not assign many TTV species; the x-axis represents classifiers, while the y-axis lists viral species, including unclassified *Alphatorquevirus*. **D** Natural log-transformed boxplot of read counts, with DIAMOND yielding fewer reads, with outliers corresponding to TTV 16 and unclassified *Alphatorquevirus*. **E**–**G** Shannon, Simpson, and Richness diversity indices, respectively. **H** Rarefaction curves, with solid lines representing all viruses and dotted lines indicating only TTV species. **I** Same as (**H**), but with the x-axis limited to 30,000 reads for a clearer view of the initial rarefaction pattern

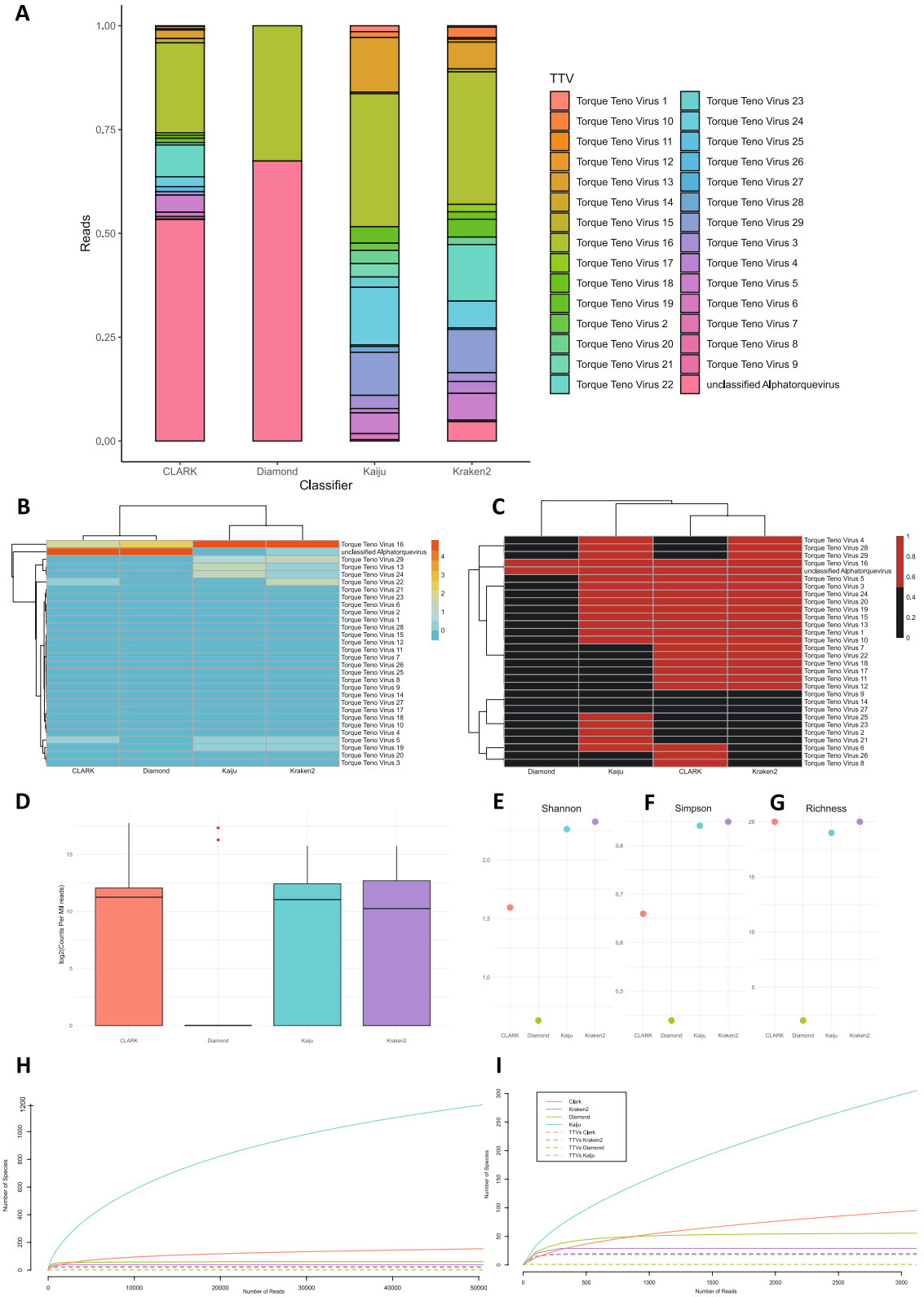

Fig. 3 (See legend on previous page.)

in DIAMOND comparisons, as visually highlighted in Fig. 3C (Table 3).

Different from what was observed for AL1, the overall viral rarefaction curve indicated that Kaiju identifies more viral species, reaching a plateau later (Fig. 3H, I). However, for TTVs specifically, all classifiers exhibited similar performance, with DIAMOND detecting fewer species, as previously noted. Although CLARK, Kaiju, and Kraken2 assigned similar TTV read counts (Fig. 3D, H, Table 3), CLARK detected more species, indicating a potentially greater sensitivity. This aligns fully with the behavior observed for the sample AL1.

Pool PR3: plasma samples from patients with prostate cancer

According to the abundance obtained by Zanette et al., 2023, in this sample, a predominant finding was attributed to Mastadenovirus, Human Pegivirus-1, and TTV. Regarding the anelloviruses, DIAMOND attributed only a few species to these, and most of the reads were attributed as unclassified alphatorqueviruses (Fig. 4A, D), showing the lowest richness index (Fig. 4G). Kraken2 and CLARK identified the highest number of species, with Kraken2 reaching the higher Shannon index (2.6 - Fig. 4E). Kaiju performed similarly in both previous pools, several TTV species, and with a high number, which is shown in Fig. 4A, B.

Considering the number of reads assigned to each TTV species, Kaiju and Kraken2 exhibited similar performance, as they clustered together. This was followed by a separate cluster comprising CLARK and DIAMOND (Fig. 4B). However, as previously mentioned, CLARK and Kraken2 identified a higher number of TTV species, clustering together in the binary heatmap (Fig. 4C). They were subsequently grouped with Kaiju, while DIAMOND formed the outermost group. Despite showing a similar number of reads, the Mann–Whitney test revealed significant differences in read counts between Kraken2 and Kaiju, as well as between CLARK and Kaiju. Additionally, all classifiers were statistically different from DIAMOND due to its low number of TTV readings (Table 4).

Based on the rarefaction curve, the TTV species identified were very similar among Kraken2, Kaiju, and CLARK (Fig. 4H, I), with DIAMOND identifying the lowest diversity (Fig. 4A). However, in a broader context, DIAMOND identified more viral species than Kraken2. Kaiju reached a plateau with approximately 200 viral species, while CLARK has not yet reached a state of rarefaction.

Table 3 Mann–Whitney Test results for PR2

Classifier 1	Classifier 2	*p*-value-adjusted*
Kraken 2	Kaiju	9.578743e−01
Kraken 2	DIAMOND	5.809736e−05
Kraken 2	CLARK	9.578743e−01
Kaiju	DIAMOND	8.659291e−05
Kaiju	CLARK	9.578743e−01
DIAMOND	CLARK	5.809736e−05

* *p*-value adjusted using BH method

Discussion

mNGS has revolutionized virus identification and discovery, driving significant advancements in bioinformatics tools for taxonomic classification. However, the choice of the algorithm can influence the obtained results, which can significantly impact the interpretation. Short-read classifiers are widely used for viral metagenomics due to enhanced coverage and sequencing depths provided by this technology. In this context, we evaluated the performance of four taxonomic classifiers in their ability to classify TTV types, which are universally present across diverse clinical samples. The abundance can be regarded as a suitable marker for assessing classifier performance while also highlighting their potential application in virus identification within clinical settings.

TTVs, belonging to the *Alphatorquevirus* genus of the *Anelloviridae* family are the most abundant viruses detected in metagenomic analysis. These commensal viruses have not been associated with any clinically relevant symptoms to date. Similarly, they are very widely distributed worldwide, with a prevalence among the population studies reporting rates of up to 99% [37–39]. In addition to their prevalence, TTVs demonstrate remarkable genetic diversity, with 26 species featuring complete genomes currently available in the NCBI RefSeq

(See figure on next page.)

Fig. 4 Comparison of TTV abundance and diversity across different taxonomic classifiers for PR3. **A** Relative abundance of TTVs identified by each classifier, with relative abundance values from 0 to 1.0 on the y-axis. **B** Z-score normalized heatmap of read counts, showing two main clusters: one formed by Kaiju and Kraken2, and another by CLARK and DIAMOND. **C** Qualitative heatmap of the detected species, DIAMOND forming the outermost cluster due to its limited assignment of TTV species; the x-axis represents classifiers, while the y-axis lists viral species, including unclassified *Alphatorquevirus*. **D** Natural log-transformed boxplot of read counts, with DIAMOND yielding the fewest reads, with outliers corresponding to TTVs 5, 15, 16, 29, and unclassified *Alphatorquevirus*. **E**–**G** Shannon, Simpson, and Richness diversity indices, respectively. **H** Rarefaction curves, with solid lines representing all viruses and dotted lines indicating only TTV species. **I** Same as (**H**), but with the x-axis limited to 30,000 reads for a clearer view of the initial rarefaction pattern

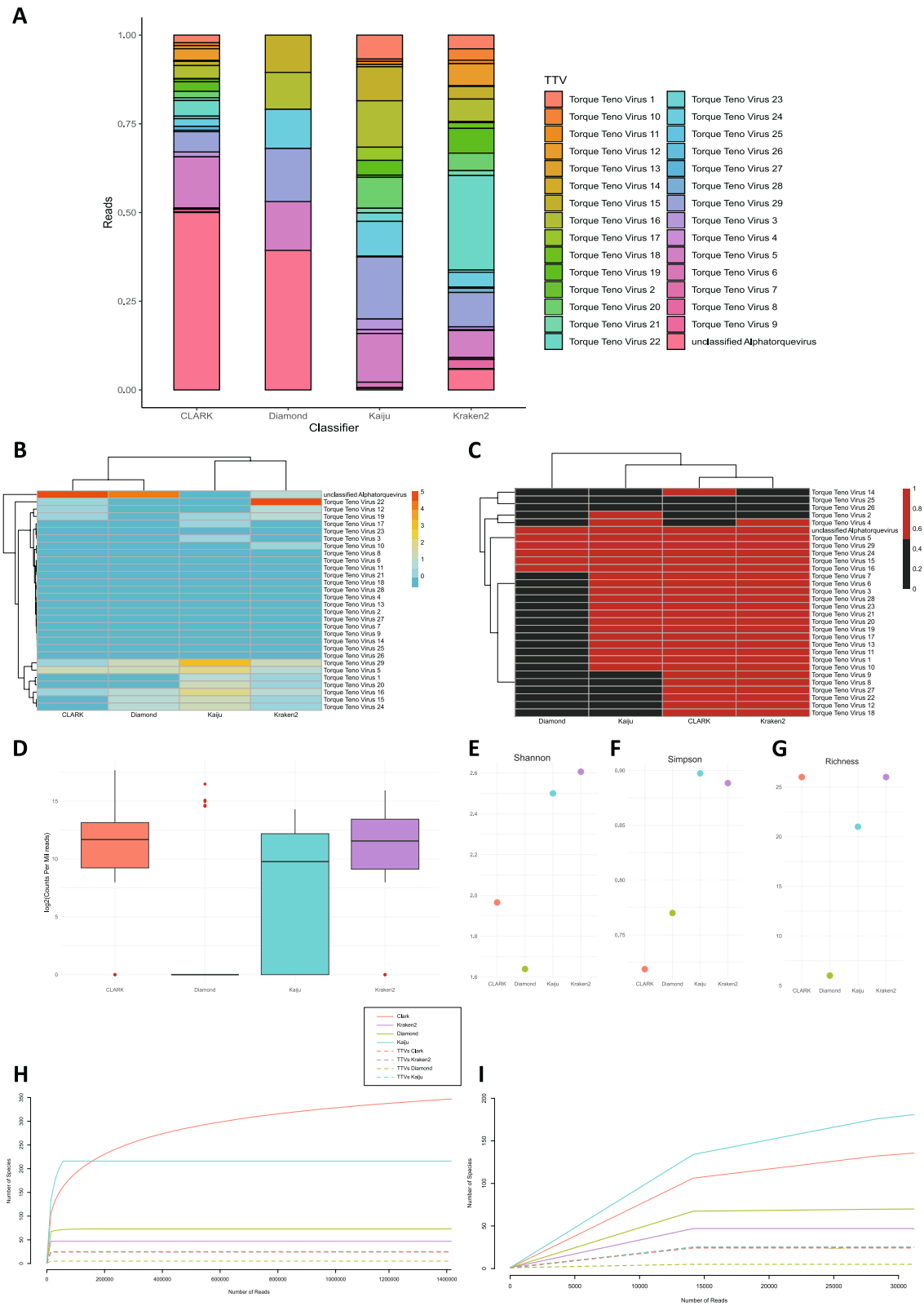

Fig. 4 (See legend on previous page.)

Table 4 Mann–Whitney Test results for PR3

Classifier 1	Classifier 2	*p*-value-adjusted*
Kraken 2	Kaiju	1.064329e−01
Kraken 2	DIAMOND	1.232224e−03
Kraken 2	CLARK	9.468548e−01
Kaiju	DIAMOND	1.783657e−02
Kaiju	CLARK	1.063077e−01
DIAMOND	CLARK	1.232224e−03

* *p*-value adjusted using BH method

database. However, other studies suggest that their true diversity is greater, with new putative subspecies being continuously identified [17, 40]. This extensive prevalence and diversity underscore their suitability as a potential marker for taxonomic classifiers accurately identify and adequately show the viral abundance.

We observed distinct performance variations across all taxonomic classifiers, with differences in the total viral reads and TTV abundance assigned by each. The lowest TTV diversity was observed for the protein-based classifier DIAMOND. In this relation, the classified TTVs were the most external group in the generated heatmaps for samples PR2 and PR3 (Fig. 3C, 4C). We hypothesized that this outcome may be related to the lower diversity indexes observed for these samples, where a substantial number of TTV reads were detected but classified into a limited number of species. The taxonomic classifier DIAMOND is based on double-indexing which can explain the reached low abundance, as such a strategy limits false-positive classification [25]. We believe that another contributing factor may be related to the capsid protein similarity among TTVs. Although at the genomic level, these viruses are highly diverse, their structural proteins are homologous, which might have contributed to the lower abundance observed by the DIAMOND classifier [39, 41]. Additionally, according to Carbo et al., 2022, protein-based classification can be less sensible to newly emerging viruses or be highly variable due to the lower mutation rates of the proteins, including the structural viral peptides [42, 43].

The other analyzed protein-based classifier, Kaiju, demonstrated distinct performance compared to DIAMOND, detecting a higher TTV diversity and attributing a relatively small number of sequences to unclassified alphatorqueteno viruses. This enhanced diversity detection is based on Kaiju´s algorithm to classify sequences, which first searches for maximum exact matches (MEM) of amino acid sequences in the given database and then performs a greedy search [24]. This dual search strategy allows Kaiju to detect amino acid substitutions more effectively, achieving higher sensitivity and precision in classification to identify the maximum number of present TTV sequences. The classification did not assign a high number of sequences read to individual TTV species, as their relative abundances were more balanced. Additionally, the high alpha diversity indexes observed suggest that this classifier is well-suited for detecting a broad range of TTV species, albeit with a lower number of reads per species. This trend is also evident in the heatmaps (Figs. 2B, 3B, 4B). However, a fundamental limitation of protein-level sequence classification lies in its inability to classify reads from non-protein-coding regions [24]. In scenarios where microbial genomes are present in the reference database including such regions, Kaiju and DIAMOND may be less sensitive than nucleotide-level classifiers, which are better equipped to handle these sequences.

Contrasting to the results obtained by DIAMOND and Kaiju, the nucleotide-based classifier CLARK demonstrated the highest sensitivity, identifying a broad range of TTV species and yielding a consistently high richness index across all evaluated samples. CLARK was applied in its "full" mode, exhibiting greater sensitivity than the other classifiers. This full mode comprehends the value for *k*-mer length $k=20$, which maximizes the number of assigned reads [23]. This setting may explain why, across all three analyzed samples, we obtained the same value for the viral and TTV reads with rarefaction curves reaching higher values without saturation (Figs. 2H, 3H, 4H).

Despite the observed ability of CLARK to identify a large subset of TTVs species, the observed abundance may not fully represent reliable diversity. This is suggested by the low Simpson and Shannon index values, which could indicate that a higher number of TTV species were assigned with relatively few reads, reflecting an uneven diversity distribution.

Kraken2, probably the most widely used classifier in metagenomics, showed a qualitative heatmap profile similar to that of CLARK. This result was expected, as Kraken2 is also nucleotide-based and identifies a high abundance of TTV species through a similar algorithm using K-mers. Highly variable viral species belonging to the same genus, as seen in this case, can be easily misclassified using the K-mer algorithm [44]. For viral detection, Kraken2 may not be as sensitive as CLARK, but it tends to produce fewer false positives. These characteristics ensure that identified reads are more likely to be accurate, even if some lower-abundance or distantly related viral taxa are missed. This characteristic makes Kraken2 reliable, though it may be less comprehensive in detecting rare or novel viral strains. The classification potential of Kraken2 was more like that of Kaiju when considering only the TTV reads, both in terms of read numbers and diversity indexes. This

observation underscores those classifiers employing completely different algorithms—K-mer and MEM—and distinct genetic material databases (nucleotides and amino acids, respectively) can exhibit comparable results [21, 24, 44]. However, translated search classifiers may have the advantage in scenarios with high genetic variability and sparsity of available reference genomes.

Conclusion

In this study, we evaluated the performance of nucleotide- and protein-based taxonomic classifiers in identifying TTVs within metagenomic datasets derived from clinical samples across different disease contexts. Our findings highlight the distinct advantages and limitations of each classification approach. Nucleotide-based classifiers, such as CLARK and Kraken2, demonstrated superior sensitivity and are valuable for identifying a broader range of TTV species, making them particularly useful for detecting emerging or rare viruses, especially in unresolved clinical cases. In contrast, protein-based classifiers, including DIAMOND and Kaiju, demonstrate greater robustness in identifying known viral species with lower protein variability. These results underscore the importance of aligning the choice of taxonomic classifier with the novel or emerging viruses, the specific objectives of the study and the type of sample analyzed, given the unique strengths and limitations of each approach. Future research should explore the combined application of these classifiers, along with the inclusion of RNA viruses and the use of mock datasets, to enhance accuracy, precision, and reliability in metagenomic studies. By tailoring classifier selection to study goals and sample composition, researchers can optimize the utility of metagenomic sequencing for both clinical diagnostics and broader viromic investigations.

Abbreviations

TTV	Torque teno virus
mNGS	Metgenomic next-generation sequencing
DENV	Dengue virus
ZIKV	Zika virus
CHIKV	Chikungunya virus
BWA	Burrows–Wheeler Aligner
BH	Benjamini–Hochberg
CPM	Counts per mil reads
HCV	Hepatitis C virus
MEM	Maximum exact matches

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12985-025-02708-8.

Supplementary file 1.

Supplementary file 2.

Acknowledgements

We are grateful to Sandra Navarro Bresciani for the figures.

Author contributions

G.M.C.: conceptualization, data curation, formal analysis, investigation, methodology, software, validation, visualization, writing-original draft preparation, writing-review and editing; L.G.C.: data curation, formal analysis, investigation, methodology, software, validation, visualization, writing-review and editing; A.R.J.L.: formal analysis, investigation, methodology, validation, visualization; writing-review and editing; E.C.: formal analysis, investigation, software, writing-review and editing; V.F.: formal analysis, investigation, software, writing-review and editing; J.P.B.X.: formal analysis, investigation, software, writing-review and editing; M.Y.N.J.: formal analysis, investigation, software, writing-review and editing; E.C.: formal analysis, investigation, software, writing-review and editing; S.C.S.: conceptualization, funding acquisition; formal analysis, investigation, software, writing-review and editing; M.G.: formal analysis, investigation, software, writing-original draft, writing-review and editing; M.C.E.: conceptualization, funding acquisition, formal analysis, investigation, software, writing-review and editing S.N.S: conceptualization, funding acquisition, investigation, project administration, resources, supervision, validation, writing-original draft, writing-review and editing.

Funding

This project was financially supported by the São Paulo Research Foundation (FAPESP) under project numbers 2017/23205-8 and 2021/11944-6, as well as the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) under project number 403075/2023-8. SNS receives a productivity scholarship (PQ) from CNPq number 305111/2022-1. Maria Carolina Elias receives grant number 31125/2021 from CNPq. The following scholarship from FAPESP was awarded: GMC (2022/00910-6; 2023/12155-0).

Data availability

The code used in this work can be found online at https://github.com/gmcampos27/MetaviromePipeline.

Declarations

Competing interests

The authors declare no competing interests.

Ethical approval

Sample collection approving can be found in Souza JVC, Santos H de O, Leite AB, Giovanetti M, Bezerra R dos S, Carvalho E de, et al. Viral Metagenomics for the Identification of Emerging Infections in Clinical Samples with Inconclusive Dengue, Zika, and Chikungunya Viral Amplification. Viruses. 2022 Aug 31;14(9):1933 and Luciola Zanette D, Andrade Coelho KBC, de Carvalho E, Aoki MN, Nardin JM, Araújo Lalli L, et al. Metagenomic insights into the plasma virome of Brazilian patients with prostate cancer. Mol Cell Oncol. 2023 Dec 31;10(1).

Author details

[1]Programa de Pós-graduação em Oncologia Clínica, Células-Tronco e Terapia Celular, Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, Ribeirão Prêto, Brazil. [2]Escola Superior de Agricultura Luiz de Queiroz, Departamento de Zootecnia, Universidade de São Paulo, Piracicaba, Brazil. [3]Centro de Vigilância Viral e Avaliação Sorológica- CeVIVas, Instituto Butantan, São Paulo, Brazil. [4]Burnett School of Medical Sciences, College of Medicine, University of Central Florida, Orlando, FL, USA. [5]Departamento de Ciências Exatas e Terra, Universidade Estadual da Bahia, Salvador, Brazil. [6]Centre for Epidemic Response and Innovation (CERI), School of Data Science and Computational Thinking, Stellenbosch University, Stellenbosch, South Africa. [7]Departamento de Análises Clínicas, Toxicológicas e Bromatológicas, Faculdade de Ciências Farmacêuticas de Ribeirão Preto, Universidade de São Paulo, Ribeirão Prêto, Brazil. [8]Laboratório de Toxinologia Aplicada, Instituto Butantan, São Paulo, Brazil. [9]Laboratório de Bacteriologia, Instituto Butantan, São Paulo, Brazil. [10]Department of Science and Technologies for Sustainable Development and One Health, Università Campus Bio-Medico di Roma, Rome, Italy. [11]Instituto Rene Rachou, Fundação Oswaldo Cruz-FIOCRUZ, Belo Horizonte, Brazil.

Received: 31 January 2025 Accepted: 13 March 2025
Published online: 28 March 2025

References

1. Kiselev D, Matsvay A, Abramov I, Dedkov V, Shipulin G, Khafizov K. Current trends in diagnostics of viral infections of unknown etiology. Viruses. 2020;12(2):211.
2. van Rijn AL, van Boheemen S, Sidorov I, Carbo EC, Pappas N, Mei H, et al. The respiratory virome and exacerbations in patients with chronic obstructive pulmonary disease. PLoS ONE. 2019;14(10): e0223952.
3. de Vries JJC, Brown JR, Fischer N, Sidorov IA, Morfopoulou S, Huang J, et al. Benchmark of thirteen bioinformatic pipelines for metagenomic virus diagnostics using datasets from clinical samples. Journal Clin Virol. 2021;141: 104908.
4. Ogunbayo AE, Sabiu S, Nyaga MM. Evaluation of extraction and enrichment methods for recovery of respiratory RNA viruses in a metagenomics approach. J Virol Methods. 2023;314: 114677.
5. Escobar-Zepeda A, Vera-Ponce de León A, Sanchez-Flores A. The road to metagenomics: from microbiology to dna sequencing technologies and bioinformatics. Front Genet. 2015;6:348.
6. Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. Shotgun metagenomics, from sampling to analysis. Nat Biotechnol. 2017;35(9):833–44.
7. Ye SH, Siddle KJ, Park DJ, Sabeti PC. Benchmarking metagenomics tools for taxonomic classification. Cell. 2019;178(4):779–94.
8. Govender KN, Eyre DW. Benchmarking taxonomic classifiers with Illumina and Nanopore sequence data for clinical metagenomic diagnostic applications. Microb Genom. 2022. https://doi.org/10.1099/mgen.0.000886.
9. Portik DM, Brown CT, Pierce-Ward NT. Evaluation of taxonomic classification and profiling methods for long-read shotgun metagenomic sequencing datasets. BMC Bioinform. 2022;23(1):541.
10. Carbo E, Sidorov I, van Rijn-Klink A, Pappas N, van Boheemen S, Mei H, et al. Performance of five metagenomic classifiers for virus pathogen detection using respiratory samples from a clinical cohort. Pathogens. 2022;11(3):340.
11. Wu LY, Wijesekara Y, Piedade GJ, Pappas N, Brussaard CPD, Dutilh BE. Benchmarking bioinformatic virus identification tools using real-world metagenomic data across biomes. Genome Biol. 2024;25(1):97.
12. Nooij S, Schmitz D, Vennema H, Kroneman A, Koopmans MPG. Overview of virus metagenomic classification methods and their biological applications. Front Microbiol. 2018;23:9.
13. Buffet-Bataillon S, Rizk G, Cattoir V, Sassi M, Thibault V, Del Giudice J, et al. Efficient and quality-optimized metagenomic pipeline designed for taxonomic classification in routine microbiological clinical tests. Microorganisms. 2022;10(4):711.
14. Cao L, Ma Y, Wan Z, Li B, Tian W, Zhang C, et al. Longitudinal anellome dynamics in the upper respiratory tract of children with acute respiratory tract infections. Virus Evol. 2023. https://doi.org/10.1093/ve/vead045.
15. Ullah Khan N, Sadiq A, Khan J, Basharat N, Hassan ZU, Ali I, et al. Molecular characterization of plasma virome of hepatocellular carcinoma (HCC) patients. AMB Express. 2024;14(1):46.
16. Dodi G, Attanasi M, Di Filippo P, Di Pillo S, Chiarelli F. Virome in the lungs: the role of anelloviruses in childhood respiratory diseases. Microorganisms. 2021;9(7):1357.
17. Cebriá-Mendoza M, Arbona C, Larrea L, Díaz W, Arnau V, Peña C, et al. Deep viral blood metagenomics reveals extensive anellovirus diversity in healthy humans. Sci Rep. 2021;11(1):6921.
18. Slavov SN, Maraninchi Silveira R, Hespanhol MR, Sauvage V, Rodrigues ES, Fontanari Krause L, et al. Human pegivirus-1 (HPgV-1) RNA prevalence and genotypes in volunteer blood donors from the Brazilian Amazon. Transfus Clin Biol. 2019;26(4):234–9.
19. Boodram B, Hershow RC, Klinzman D, Stapleton JT. GB virus C infection among young, HIV-negative injection drug users with and without hepatitis C virus infection. J Viral Hepat. 2011. https://doi.org/10.1111/j.1365-2893.2010.01350.x.
20. Aziz RK, Dwivedi B, Akhter S, Breitbart M, Edwards RA. Multidimensional metrics for estimating phage abundance, distribution, gene density, and sequence coverage in metagenomes. Front Microbiol. 2015;8:6.
21. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol. 2019;20:1–13.
22. Lawrence JG, Hatfull GF, Hendrix RW. Imbroglios of viral taxonomy: genetic exchange and failings of phenetic approaches. J Bacteriol. 2002;184(17):4891–905.
23. Ounit R, Wanamaker S, Close TJ, Lonardi S. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. BMC Genomics. 2015. https://doi.org/10.1186/s12864-015-1419-2.
24. Menzel P, Ng KL, Krogh A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nat Commun. 2016;7(1):11257.
25. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12(1):59–60.
26. Souza JVC, de Santos HO, Leite AB, Giovanetti M, dos Bezerra RS, de Carvalho E, et al. Viral Metagenomics for the Identification of Emerging Infections in Clinical Samples with Inconclusive Dengue, Zika, and Chikungunya Viral Amplification. Viruses. 2022;14(9):1933.
27. Luciola Zanette D, Andrade Coelho KBC, de Carvalho E, Aoki MN, Nardin JM, Araújo Lalli L, et al. Metagenomic insights into the plasma virome of Brazilian patients with prostate cancer. Mol Cell Oncol. 2023. https://doi.org/10.1080/23723556.2023.2188858.
28. de Iani FCM, de Campos GM, Adelino TER, da Silva AS, Kashima S, Alcantara LCJ, et al. Metagenomic analysis for diagnosis of hemorrhagic fever in Minas Gerais, Brazil. Microorganisms. 2024;12(4):769.
29. Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc
30. Chen S, Zhou Y, Chen Y, Gu J. Fastp: An ultra-fast all-in-one FASTQ preprocessor. In: Bioinformatics. 2018
31. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–60.
32. Bağcı C, Patz S, Huson DH. DIAMOND+MEGAN: fast and easy taxonomic and functional analysis of short and long microbiome sequences. Curr Protoc. 2021. https://doi.org/10.1002/cpz1.59.
33. R Core Team. A Language and Environment for Statistical Computing_. R Foundation for Statistical Computing, 2023
34. Oksanen J, Simpson GL, Blanchet FG, Kindt R, Legendre P, Minchin PR, et al. vegan: Community Ecology Package [Internet]. 2024. Available from: https://CRAN.R-project.org/package=vegan
35. Wickham H. ggplot2: Elegant Graphics for Data Analysis [Internet]. Springer-Verlag New York; 2016. Available from: https://ggplot2.tidyverse.org
36. Ram K, Wickham H. wesanderson: A Wes Anderson Palette Generator [Internet]. 2023. Available from: https://CRAN.R-project.org/package=wesanderson
37. Kaczorowska J, Deijs M, Klein M, Bakker M, Jebbink MF, Sparreboom M, et al. Diversity and long-term dynamics of human blood anelloviruses. J Virol. 2022. https://doi.org/10.1128/jvi.00109-22.
38. Kaczorowska J, van der Hoek L. Human anelloviruses: diverse, omnipresent and commensal members of the virome. FEMS Microbiol Rev. 2020;44(3):305–13.
39. Arze CA, Springer S, Dudas G, Patel S, Bhattacharyya A, Swaminathan H, et al. Global genome analysis reveals a vast and dynamic anellovirus landscape within the human virome. Cell Host Microbe. 2021;29(8):1305-1315.e6.
40. Kaczorowska J, Timmerman AL, Deijs M, Kinsella CM, Bakker M, van der Hoek L. Anellovirus evolution during long-term chronic infection. Virus Evol. 2023;9(1):vead001.
41. Redondo N, Navarro D, Aguado JM, Fernández-Ruiz M. Viruses, friends, and foes: the case of torque teno virus and the net state of immunosuppression. Transp Infect Dis. 2022. https://doi.org/10.1111/tid.13778.
42. Duffy S, Shackelton LA, Holmes EC. Rates of evolutionary change in viruses: patterns and determinants. Nat Rev Genet. 2008;9(4):267–76.
43. Sanjuán R, Domingo-Calap P. Mechanisms of viral mutation. Cell Mol Life Sci. 2016;73(23):4433–48.
44. Tovo A, Menzel P, Krogh A, Cosentino Lagomarsino M, Suweis S. Taxonomic classification method for metagenomics based on core protein families with Core-Kaiju. Nucleic Acids Res. 2020;48(16):e93–e93.

Publisher's Note