Virus Research 263 (2019) 21–26

Contents lists available at ScienceDirect

Virus Research

journal homepage: www.elsevier.com/locate/virusres

Enhanced bioinformatic profiling of VIDISCA libraries for virus detection and discovery

Cormac M. Kinsella*, Martin Deijs, Lia van der Hoek

Laboratory of Experimental Virology, Department of Medical Microbiology, Amsterdam UMC, University of Amsterdam, Meibergdreef 9, Amsterdam, the Netherlands

ARTICLE INFO

Keywords:
Virus discovery
Bioinformatics
Cluster-profiling
VIDISCA

ABSTRACT

VIDISCA is a next-generation sequencing (NGS) library preparation method designed to enrich viral nucleic acids from samples before highly-multiplexed low depth sequencing. Reliable detection of known viruses and discovery of novel divergent viruses from NGS data require dedicated analysis tools that are both sensitive and accurate. Existing software was utilised to design a new bioinformatic workflow for high-throughput detection and discovery of viruses from VIDISCA data. The workflow leverages the VIDISCA library preparation molecular biology, specifically the use of Mse1 restriction enzyme which produces biological replicate library inserts from identical genomes. The workflow performs total metagenomic analysis for classification of non-viral sequence including parasites and host, and separately carries out virus specific analyses. Ribosomal RNA sequence is removed to increase downstream analysis speed and remaining reads are clustered at 100% identity. Known and novel viruses are sensitively detected via alignment to a virus-only protein database, and false positives are removed. A new cluster-profiling analysis takes advantage of the viral biological replicates produced by Mse1 digestion, using read clustering to flag the presence of short genomes at very high copy number. Importantly, this analysis ensures that highly repeated sequences are identified even if no homology is detected, as is shown here with the detection of a novel gokushovirus genome from human faecal matter. The workflow was validated using read data derived from serum and faeces samples taken from HIV-1 positive adults, and serum samples from pigs that were infected with atypical porcine pestivirus.

1. Introduction

The host range expansion of viral pathogens and emergence of novel species can pose substantial threats to human health (Parrish et al., 2008). Viruses evolve rapidly, possess high molecular diversity, and are found in relatively low concentration alongside host nucleic acids in most sample types. These factors complicate detection of novel viral genetic material and necessitate specific virus discovery methods to achieve sufficient detection sensitivity. Next-generation sequencing (NGS) and metagenomics have greatly accelerated the discovery of novel viruses when contrasted with traditional wet-lab virological techniques such as isolation in cell culture, as they can be performed on any virus directly from biological or environmental samples, in a high-throughput way (Shi et al., 2018, 2016). Approaches that prioritise an unbiased metagenomic profile require high sequencing depth to ensure pathogen detection, and are therefore relatively expensive per viral nucleotide. The incorporation of virus enrichment techniques prior to sequencing reduces the required depth for detection (Conceição-Neto et al., 2015; de Vries et al., 2011), and may be desirable when processing tens to hundreds of samples.

VIDISCA is a virus discovery NGS library preparation method that enriches viral nucleic acids in samples before low depth Ion Torrent sequencing, allowing processing of 140 samples per week. The wet-lab procedure, described in detail elsewhere (de Vries et al., 2011; Edridge et al., 2018), is summarised here in order to highlight advantages for bioinformatic analysis. First, cells and debris are pelleted, and virus-containing supernatant is DNase treated to reduce residual cellular DNA. Virion proteins are linearised to release nucleic acid, which is extracted using the Boom method (Boom et al., 1990). RNA viruses are reverse transcribed using non-ribosomal RNA (rRNA) hexamer primers (Endoh et al., 2005), which reduce the proportion of rRNA transcribed into DNA. After second-strand synthesis, double-stranded DNA products are digested using the frequent cutting Mse1 restriction enzyme, an important feature unique to VIDISCA library preparation. Sequencing primers are ligated onto the two sticky ends of a restriction fragment, before size selection against both long and short fragments, amplification with PCR, and sequencing with the Ion Torrent PGM platform (Thermo Fisher Scientific, Waltham, MA, USA).

* Corresponding author.
E-mail address: c.m.kinsella@amc.uva.nl (C.M. Kinsella).

https://doi.org/10.1016/j.virusres.2018.12.010
Received 1 September 2018; Received in revised form 8 December 2018; Accepted 18 December 2018
Available online 19 December 2018

The inclusion of Mse1 digestion during library preparation has advantageous implications for virus discovery bioinformatics. Viral genomes are short compared to their host, and can be at high copy number during infection. Since Mse1 reproducibly cuts homologous restriction fragments from genomes of the same type, high numbers of viral biological replicates with identical start and end sites are expected in library inserts prior to PCR. This is in contrast with a randomly fragmented library in which identical start and end sites are relatively rare. The VIDISCA insert redundancy is not expected from background or host nucleic acid, except that with 'virus-like' characteristics, i.e. high copy number, such as mitochondrial DNA. The virus replicates should result in characteristic redundancy in sequencing data, which can be identified via read clustering. Additionally, since Mse1 cuts TTAA sites, it cuts more rarely in GC rich rRNA (de Vries et al., 2011). Viable rRNA VIDISCA fragments are generally longer as a result, and can be disproportionately reduced during size selection, contributing to a high sensitivity that enables lower sequencing depth and analysis time. Recently VIDISCA was used to discover the suspected human pathogen Ntwetwe virus with 2 reads from 6,947, whereas an in-house Illumina workflow optimised for virus detection found only 8 reads among the 2,741,915 obtained (Edridge et al., 2018).

Here we present a new bioinformatic workflow designed to process VIDISCA data. The core task is sensitive virus detection including false positive reduction. The workflow includes metagenomic analysis for identification of host background and non-viral organisms including parasites, and collects descriptive metrics in order to flag unusual properties of samples, such as high rRNA content. It outputs text and interactive HTML results for detailed investigation of samples, and includes a new cluster-profiling analysis used to flag the presence of sequences at high copy number (e.g. virus infections). This analysis also provides an informative profile of sample content in different classification bins, including known and novel viruses, mitochondrial DNA, and background sequence. Notably, the flagging of highly repetitive reads does not rely on identity searches, ensuring that abundant unknown sequences can be identified. The utility of the workflow is presented with examples.

2. Materials and methods

2.1. Bioinformatic workflow for VIDISCA next-generation sequencing data

The new bioinformatic workflow for VIDISCA NGS data is summarised graphically (Fig. 1) and described in detail below. As input, the workflow takes FASTA formatted sequences. Eukaryotic and prokaryotic virus protein databases used by the workflow were constructed in advance from respective NCBI Identical Protein Groups datasets, followed by clustering at 95% identity using CD-HIT v4.7 (Fu et al., 2012). First, metagenomic analysis of raw reads is carried out using Centrifuge v1.0.3 (Kim et al., 2016) against the pre-built NCBI non-redundant nucleotide Centrifuge index including known viruses, eukaryotes, and prokaryotes (February 2018). Centrifuge classification tables are visualised as interactive HTML charts using Recentrifuge (Martí, 2018).

Next, the main virus detection steps are run. Reads from rRNA are separated from raw reads using SortMeRNA v2.1 (Kopylova et al., 2012). Non-rRNA reads are sorted by length and clustered at 100% identity using CD-HIT v4.7, and 'clstr' files are retained for later processing. Clustered non-rRNA reads are queried against the eukaryotic virus protein database using the UBLAST algorithm provided as part of the USEARCH v10 software package, with -mincodons set to 15, -accel to 0.8, and -evalue to 1e-4 (Edgar, 2010). Unmatched reads from this step are queried against the prokaryotic virus protein database, and those remaining unclassified are mapped to human, pig, and chicken mitochondrial DNA sequences using the BWA-MEM algorithm of BWA v0.7.17 (Li, 2013). Reads matching the eukaryotic virus protein database are treated as putatively viral, and are next queried against the NCBI nt. database (April 2018) using BLASTn v2.4.0 (Camacho et al., 2009). Those classified by BLASTn as viral are regarded as confident viral reads (classified as viral twice), those classified as non-viral are regarded as false positives, and those that remain unclassified are regarded as possible unknown viruses (classified as viral once). This information is used to split the UBLAST protein classification tables into the three categories, each of which are visualised separately as interactive HTML charts using KronaTools v2.7 (Ondov et al., 2011). The BLASTn classification of false positives is also visualised for inspection and comparison to the original viral classification.

Cluster-profiling outputs are produced using the CD-HIT 'clstr' files, which are converted into a table reporting the representative sequences, the number of reads clustered per representative, and the proportion of the original non-rRNA that each represents in a sample. The classification bin (such as 'confident virus', or 'mitochondrial DNA') of each representative read is then added to the table, including a bin for unclassified sequences. This output is plotted as a bar chart using ggplot2, with separate bars for classification bins, and representative reads stacked according to proportional amount of clustering (Wickham, 2016). The classification bins are 'Virus (aa + nt)' including reads classified as viral twice, 'Virus (aa)' including reads classified as viral once, 'False pos. (nt)' including reads removed as probable false positives, 'Phage (aa)' including reads aligning to our prokaryotic virus database, 'MitoDNA' including reads mapped to mitochondrial DNA references, 'Centrifuge' including reads identified by the metagenomic tool Centrifuge, and 'No hit' including reads with no assigned classification. The bar chart output provides a visual overview of the proportion of reads from a sample that were classified in a particular bin. Furthermore, reads that represent many other reads are visually

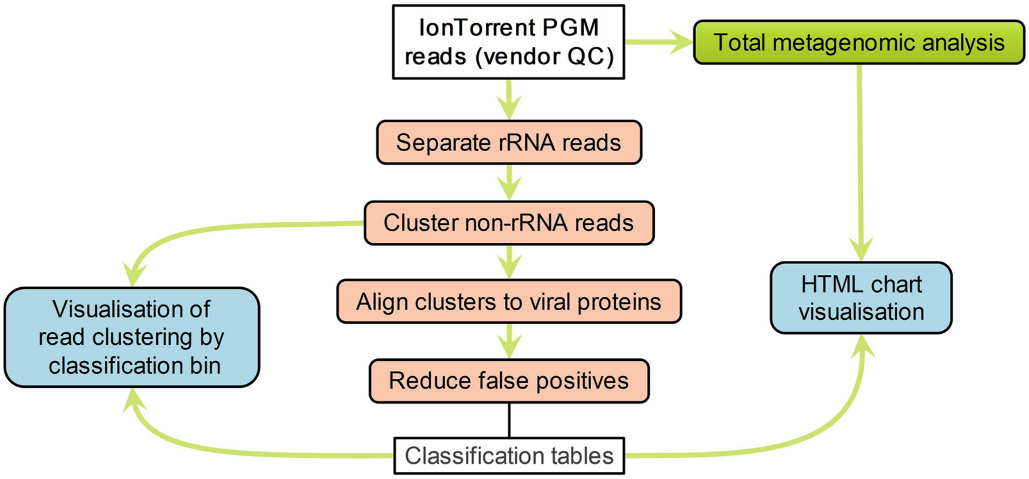
IonTorrent PGM reads (vendor QC)
Total metagenomic analysis
Separate rRNA reads
Cluster non-rRNA reads
Align clusters to viral proteins
Reduce false positives
Classification tables
Visualisation of read clustering by classification bin
HTML chart visualisation

Fig. 1. Schematic overview of the bioinformatic workflow for VIDISCA data, showing the main virus detection and discovery steps (orange), the metagenomic analysis (green), and visualisation processes (blue). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

identifiable due to their higher relative proportion. This allows the presence of clustering to be identified in each bin separately. Most repetitive non-viral sequences are accounted for via removal of rRNA and binning of mitochondrial DNA, however unclassified sequences putatively from viruses require manual inspection or full-length sequencing in order to establish their likely provenance.

For each classification bin, the 10 representative sequences accounting for the largest proportion of reads are automatically extracted as FASTA files for inspection, for example with BLASTx. All text tables and sample-specific files produced by the analysis are packaged into sample folders, and descriptive metrics about the run time and classification performance for each sample are reported to a log file for later examination.

2.2. Data selection and workflow testing

Three VIDISCA datasets were selected and analysed using the new bioinformatic workflow, in order to assess specific aspects of workflow performance and utility. First, VIDISCA reads from 194 serum samples collected in 1994–1995 from HIV-1 infected adults were run. The aim was to determine whether the bioinformatic workflow outputs could be used to troubleshoot the likely causes of pathogen detection failure. This was done by comparison of HIV-1 detection by VIDISCA with preexisting HIV-1 load data obtained using nucleic acid sequence based amplification (NASBA). Outputs from samples in which HIV-1 was unexpectedly not detected were manually inspected to determine the cause of failure.

Second, VIDISCA reads from 194 faecal samples from the above mentioned cohort were run (Oude Munnink et al., 2014). The aim was to test the prediction that cluster-profiling could be used to flag virus-like characteristics in unclassified reads, and therefore identify novel viruses at high load missed by classification algorithms. Cluster-profiling outputs were examined for evidence of clustering among unclassified reads and a single sample (F115) was selected for follow up. Illumina reads from a randomly fragmented library of the sample were downloaded from the European Nucleotide Archive (accession ERR233419), cleaned of adapters, quality trimmed (minimum 50bp, sliding window trim < Q20) with Trimmomatic v0.38 (Bolger et al., 2014), and assembled using SPAdes v3.12 (Bankevich et al., 2012). The 10 unclassified VIDISCA representative sequences accounting for the most clustering were BLAST queried against the contigs, and the most common target sequence was extracted and manually curated.

Third, VIDISCA reads from 13 serum samples taken from sows experimentally infected with atypical porcine pestivirus (APPV) and 16 serum samples taken from the transplacentally-infected piglets of the sows were run (de Groof et al., 2016). In this case, sequencing was carried out on an Ion Proton instrument (Thermo Fisher Scientific, Waltham, MA, USA). The aims were to statistically test support for the assumption that a higher viral load would result in higher clustering among viral reads, and to explore whether such an association was strongly influenced by PCR bias toward abundant templates. Since the dataset included individuals infected with the same virus strain at a large range of viral loads, this was carried out as a reliability test of the main assumption underlying cluster-profiling analysis, that VIDISCA library preparation selects for biological replicates from identical genomes, resulting in read clustering associated with the biological load of a sequence.

3. Results and discussion

3.1. Bioinformatic workflow design

The new VIDISCA bioinformatic workflow has been designed to prioritise sensitivity to viruses, however non-virus metagenomics and the efficiency of analysis have also been considered. *K*-mer based metagenomic tools such as Kraken (Wood and Salzberg, 2014) are commonly used for pathogen detection, since they provide very rapid classification of reads via exact matches of length *k* between reads and reference indexes. Metagenomic samples often contain species with variable nucleotide identity to their most related reference sequence. Since *k* must be set in advance, high *k* decreases classification sensitivity for distantly related species, and low *k* decreases precision to well represented taxa. To circumvent this, the metagenomic software tool Centrifuge was selected for the workflow since it uses FM-indexed reference sequences, allowing *k* to be optimal for each individual read in a metagenomic sample, maximising both sensitivity and precision while simultaneously minimising index size and memory requirements (Kim et al., 2016).

Detection of novel viruses is normally achieved via local alignment of reads to viral proteins, a computationally intensive operation. High speed algorithms are available to decrease analysis time, for example UBLAST (Edgar, 2010), DIAMOND (Buchfink et al., 2015), or Kaiju (Menzel et al., 2016). Minimisation of query reads and database size can provide additional gains. The VIDISCA workflow incorporates several of these speed-ups, including rRNA removal to reduce query reads, and redundancy removal in non-rRNA using clustering. Clustering information is retained for retrospective classification of redundant reads and cluster-profiling analysis. These steps reduced average protein query counts by 31% and 45% in the 194 faecal and 194 serum datasets respectively. A virus-only protein database was constructed and clustered for a size reduction of 81%. Alignment of reads to a taxonomically restricted database raises the likelihood of spurious hits due to chance similarity, therefore false positive removal via BLAST analysis against the NCBI nucleotide database is required. Due to the prior selection steps mentioned above, a minority of reads require this querying, for example an average of 1.5% and 2.4% of reads from the above faecal and serum datasets were queried.

3.2. Assessment of the bioinformatic workflow performance

The VIDISCA bioinformatic workflow was used to identify the causes of HIV-1 detection failure in data generated from archival serum samples collected from HIV-1 positive adults. Bioinformatic analysis detected the pathogen in 128 of 194 samples (66%) with an average of 42,124 total reads per sample. Of the VIDISCA negative samples, 23 (35%) had undetectable HIV-1 loads when specifically tested with NASBA, while 9 (7%) VIDISCA positive samples did. There was a median value of 84 HIV-1 copies/μl in VIDISCA positive samples and 14 in negative (Fig. 2A), suggesting detection failure was mostly attributable to viral load. Viral load was positively associated with the proportion of HIV-1 reads (Spearman's rho = 0.61, $p < .001$), however the variance was poorly described by a linear regression model (Fig. 2B), showing that sample dependent factors crucially impact the metagenomic profile. Notably, rRNA proportion was weakly but positively associated with HIV-1 proportion (Spearman's rho = 0.34, $p < .001$), while the proportion of non-rRNA identified as human (including residual genomic DNA and cellular RNA) was found to have a weak negative association with the HIV-1 proportion (Spearman's rho = -0.17, p = .017). Together these observations imply sample-specific biases against integrity or representation of the RNA fraction. Contributing factors could include higher degradation susceptibility during freeze-thaw cycles, high host DNA content with only partial degradation during DNase treatment, high intrinsic RNase activity in certain samples, or sample-specific inhibition of reverse transcription. An additional explanation could be that rRNA acts as a carrier for low concentrations of viral RNA.

HIV-1 was not detected in 11 outlier samples with over 50 HIV-1 copies/μl and an average read count of 40,290. In 3 of these, cluster-profiling showed that 78–90% of processed (non-rRNA) reads belonged to Hepatitis B virus, which commonly dominates VIDISCA metagenomic profiles if present. One sample also showed possible competition with Torque Teno virus which represented 30% of processed reads. A

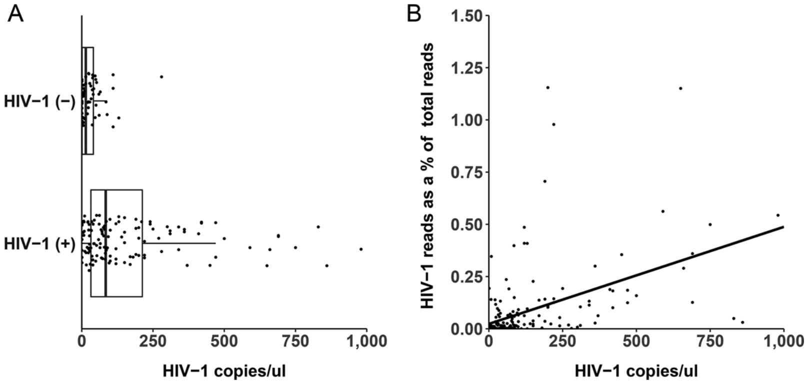

Fig. 2. A: HIV-1 viral RNA load in serum and VIDISCA outcome. HIV-1 detection in sequence reads is indicated with HIV-1 (+), and lack of detection is indication with HIV-1 (-). On the x-axis the HIV-1 RNA load per µl of serum is plotted. B: Linear regression model fitted to HIV-1 viral load against HIV-1 reads as a percentage of total reads, F(1,192) = 56.68, $p < .001$, $R^2 = 0.228$. A low 23% of variance in proportion is explained by viral load when assuming a linear relationship.

further 6 samples had approximately 80–95% of processed reads classified by Centrifuge as host or bacterial sequence with very low read clustering, suggesting a highly diverse library insert distribution probably derived from cell lysis. In the final sample an unusually high 75% of processed reads were not classified by any analysis. Manual BLAST analysis on some of these unclassified reads gave bacterial hits or weak alignment scores suspected to originate from unknown bacteriophages, suggesting bacterial growth in the stored material.

3.3. Cluster-profiling for virus discovery

A cluster-profiling analysis was incorporated in the workflow based on the prediction that short viral genomes at high load would result in distinctive read clustering characteristics, since VIDISCA library preparation produces homologous library inserts from each genome based on its Mse1 restriction sites. The analysis uses read clustering and classification information generated as part of the workflow to generate a visual output, and therefore does not require significant additional computational time. Importantly, the clustering signal generated by high copy number sequences does not require identity-based classification. This could potentially allow detection of highly divergent viruses with low protein identity to relatives represented in databases.

Cluster-profiling images generated using VIDISCA data from 194 faecal samples were analysed and sample F115 was selected for follow-up due to a high degree of clustering among unclassified reads – 12% of the 16,160 processed reads were clustered into only 100 unclassified representative sequences (Fig. 3), suggesting an unknown entity at high copy number. Available Illumina data from a randomly fragmented library of this sample were assembled into 9157 contigs. Ten unclassified representative VIDISCA sequences accounting for the most reads, which were automatically extracted by the workflow, were aligned to the contigs using BLAST. Of the 10, 8 aligned to a single contig, suggesting that they were part of a genome of a novel virus present at high copy number. Manual curation of this 5 kb sequence showed that it is a novel gokushovirus (circular ssDNA bacteriophage, NCBI accession number MK263179) with 72% nucleotide identity to its closest relative. The sequences of this virus were not identified by the classification components of the workflow since the related viral proteins were not part of the reference set. Mapping of complete read-sets revealed that 6.83% of Illumina read-pairs from the sample were derived from the virus and 17.27% of VIDISCA reads were. The result confirms the expectation that viruses at high load produce characteristic clusters in VIDISCA data, ensuring that those missed by identity searches can still be detected.

3.4. Association between viral read clustering and viral load

Cluster-profiling analysis for discovery of viruses, as shown in Fig. 3, relies on a high level of sequence redundancy in order to generate a visible signal that can be investigated. A strong association between viral load and the level of clustering observed in viral reads is expected, an effect that would underlie application of the analysis to the discovery of novel viruses. To test this assumption VIDISCA reads from 29 serum samples taken from pigs infected with APPV were analysed. The workflow detected APPV reads in 27 of these, and a strong linear association between viral load and the proportion of APPV reads was observed after removal of a single outlier (linear regression, F (1,26) = 70.57, $p < .001$, $R^2 = 0.73$). As expected, there was a strong association between viral load and the average number of reads clustered per APPV representative sequence (Spearman's rho = 0.81, $p < .001$). To account for the possibility that this effect was due to stochastic PCR bias disproportionately amplifying abundant templates (Kebschull and Zador, 2015), an association between viral load and the proportion of all APPV reads that were represented by the top APPV sequence cluster was tested for. Since viral load should correspond to the abundance of replicate templates prior to PCR, PCR bias would be expected to occur in samples with the highest loads. No such relationship existed (Spearman's rho = 0.17, p = 0.41).

Together the observations show that the degree of clustering among viral reads corresponds well with true biological load, and does not suffer from significant PCR bias toward abundant templates. While the analysis therefore can be applied to detection of novel viruses in unclassified reads, it is important to note that only infections with a high load and a high proportional amount of reads are likely to be observed. For example, it is unlikely that the analysis would have successfully flagged the presence of HIV-1 reads in the human serum samples analysed above, had they not been successfully classified using alignment tools. Nonetheless, it does provide an additional approach to both virus detection and the graphical representation of sample content, which are useful supplements to the more sensitive approaches utilised by the bioinformatic workflow.

3.5. Conclusions

A new bioinformatic workflow for sensitive virus detection and discovery in VIDISCA sequence data has been presented, which includes false positive removal and total metagenomic analysis. The workflow has been validated for virus detection in samples derived from individuals infected with known pathogens. The new cluster-profiling analysis, based on the VIDISCA library preparation molecular biology, has been used to flag a novel virus in unclassified reads, serving as a proof of concept for discovery of more divergent viruses.

Code is available upon request

For example outputs from the pipeline, see the GitHub repository at: https://github.com/CormacKinsella/VIDISCA-e.g.-output.

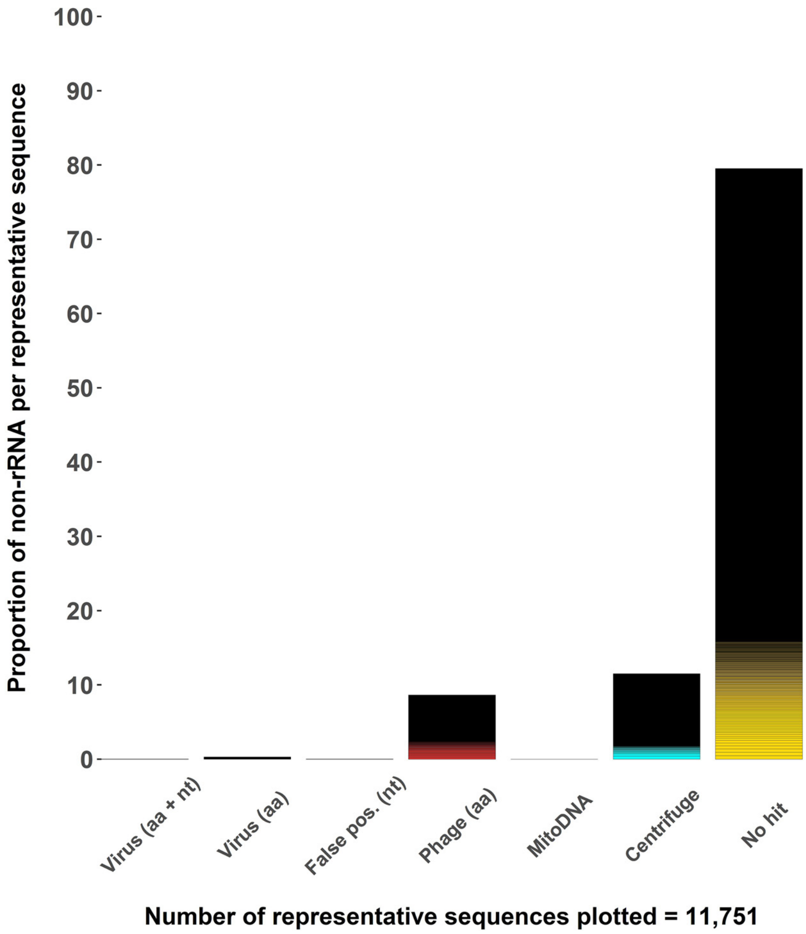

Fig. 3. Cluster-profiling bar chart from sample F115. Representative sequences produced by read clustering are plotted according to their final classification bin (x-axis) and stacked in order of their relative abundance with respect to the original non-rRNA read set (i.e. the proportion of identical reads, y-axis). Coloured bars therefore signify those sequences representing many identical reads, while many singleton reads make up black regions. Classification bins on the x-axis are those described in section 2.1. Read clustering can be seen in the phage ('Phage', red), metagenomically identified ('Centrifuge', blue), and unclassified ('No hit', yellow) read bins. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Acknowledgements

This research has received funding from the European Union's Horizon 2020 research and innovation programme, under the Marie Skłodowska-Curie Actions grant agreement no. 721367 (HONOURs). We would like to thank Dr. Ad de Groof of Intervet International BV for sharing APPV RT-qPCR data and Arthur W.D. Edridge for helpful feedback on the manuscript.

References

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., Pyshkin, A.V., Sirotkin, A.V., Vyahhi, N., Tesler, G., Alekseyev, M.A., Pevzner, P.A., 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 19, 455–477. https://doi.org/10.1089/cmb.2012.0021.

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170.

Boom, R., Sol, C.J., Salimans, M.M., Jansen, C.L., Wertheim-Van Dillen, P.M., van der Noordaa, J., 1990. Rapid and simple method for purification of nucleic acids. J. Clin. Microbiol. 28, 495–503. https://doi.org/10.1556/AMicr.58.2011.1.7.

Buchfink, B., Xie, C., Huson, D.H., 2015. Fast and sensitive protein alignment using DIAMOND. Nat. Methods 12, 59–60. https://doi.org/10.1038/nmeth.3176.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L., 2009. BLAST+: architecture and applications. BMC Bioinform. 10, 421. https://doi.org/10.1186/1471-2105-10-421.

Conceição-Neto, N., Zeller, M., Lefrère, H., De Bruyn, P., Beller, L., Deboutte, W., Yinda, C.K., Lavigne, R., Maes, P., Van Ranst, M., Heylen, E., Matthijnssens, J., 2015. Modular approach to customise sample preparation procedures for viral metagenomics: a reproducible protocol for virome analysis. Sci. Rep. 5, 16532. https://doi.org/10.1038/srep16532.

de Groof, A., Deijs, M., Guelen, L., van Grinsven, L., van Os-Galdos, L., Vogels, W., Derks, C., Cruijsen, T., Geurts, V., Vrijenhoek, M., Suijskens, J., van Doorn, P., van Leengoed, L., Schrier, C., Hoek, L., de Groof, A., Deijs, M., Guelen, L., van Grinsven, L., van Os-Galdos, L., Vogels, W., Derks, C., Cruijsen, T., Geurts, V., Vrijenhoek, M., Suijskens, J., van Doorn, P., van Leengoed, L., Schrier, C., van der Hoek, L., 2016. Atypical porcine pestivirus: a possible cause of congenital tremor type A-II in newborn piglets. Viruses 8, 271. https://doi.org/10.3390/v8100271.

de Vries, M., Deijs, M., Canuti, M., van Schaik, B.D.C., Faria, N.R., van de Garde, M.D.B., Jachimowski, L.C.M., Jebbink, M.F., Jakobs, M., Luyf, A.C.M., Coenjaerts, F.E.J., Claas, E.C.J., Molenkamp, R., Koekkoek, S.M., Lammens, C., Leus, F., Goossens, H., Ieven, M., Baas, F., van der Hoek, L., 2011. A sensitive assay for virus discovery in respiratory clinical samples. PLoS One 6, e16118. https://doi.org/10.1371/journal.pone.0016118.

Edgar, R.C., 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26, 2460–2461. https://doi.org/10.1093/bioinformatics/btq461.

Edridge, A.W.D., Deijs, M., Namazzi, R., Cristella, C., Jebbink, M.F., Maurer, I., Kootstra, N.A., Buluma, L.R., van Woensel, J.B.M., de Jong, M.D., Idro, R., Boele van Hensbroek, M., van der Hoek, L., 2018. Novel orthobunyavirus identified in the cerebrospinal fluid of a Ugandan child with severe encephalopathy. Clin. Infect. Dis. https://doi.org/10.1093/cid/ciy486.

Endoh, D., Mizutani, T., Kirisawa, R., Maki, Y., Saito, H., Kon, Y., Morikawa, S., Hayashi, M., 2005. Species-independent detection of RNA virus by representational difference analysis using non-ribosomal hexanucleotides for reverse transcription. Nucleic Acids Res. 33, e65. https://doi.org/10.1093/nar/gni064.

Fu, L., Niu, B., Zhu, Z., Wu, S., Li, W., 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28, 3150–3152. https://doi.org/10.1093/bioinformatics/bts565.

Kebschull, J.M., Zador, A.M., 2015. Sources of PCR-induced distortions in high-throughput sequencing data sets. Nucleic Acids Res. 43, e143. https://doi.org/10.1093/nar/gkv717.

Kim, D., Song, L., Breitwieser, F.P., Salzberg, S.L., 2016. Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome Res. 26, 1721–1729. https://doi.org/10.1101/gr.210641.116.

Kopylova, E., Noé, L., Touzet, H., 2012. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics 28, 3211–3217. https://doi.org/10.1093/bioinformatics/bts611.

Li, H., 2013. Aligning Sequence Reads, Clone Sequences and Assembly Contigs With BWA-MEM. arXiv:1303.3997v1 [q-bio.GN].

Martí, J.M., 2018. Recentrifuge: Robust Comparative Analysis and Contamination Removal for Metagenomic Data. bioRxivhttps://doi.org/10.1101/190934. 190934.

Menzel, P., Ng, K.L., Krogh, A., 2016. Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nat. Commun. 7, 11257. https://doi.org/10.1038/ncomms11257.

Ondov, B.D., Bergman, N.H., Phillippy, A.M., 2011. Interactive metagenomic visualization in a Web browser. BMC Bioinform. 12, 385. https://doi.org/10.1186/1471-

2105-12-385.

Oude Munnink, B.B., Canuti, M., Deijs, M., de Vries, M., Jebbink, M.F., Rebers, S., Molenkamp, R., van Hemert, F.J., Chung, K., Cotten, M., Snijders, F., Sol, C.J., van der Hoek, L., 2014. Unexplained diarrhoea in HIV-1 infected individuals. BMC Infect. Dis. 14, 22. https://doi.org/10.1186/1471-2334-14-22.

Parrish, C.R., Holmes, E.C., Morens, D.M., Park, E.-C., Burke, D.S., Calisher, C.H., Laughlin, C.A., Saif, L.J., Daszak, P., 2008. Cross-species virus transmission and the emergence of new epidemic diseases. Microbiol. Mol. Biol. Rev. 72, 457–470. https://doi.org/10.1128/MMBR.00004-08.

Shi, M., Lin, X.-D., Tian, J.-H., Chen, L.-J., Chen, X., Li, C.-X., Qin, X.-C., Li, J., Cao, J.-P., Eden, J.-S., Buchmann, J., Wang, W., Xu, J., Holmes, E.C., Zhang, Y.-Z., 2016. Redefining the invertebrate RNA virosphere. Nature 540, 1–12. https://doi.org/10.1038/nature20167.

Shi, M., Lin, X.-D., Chen, X., Tian, J.-H., Chen, L.-J., Li, K., Wang, W., Eden, J.-S., Shen, J.-J., Liu, L., Holmes, E.C., Zhang, Y.-Z., 2018. The evolutionary history of vertebrate RNA viruses. Nature 556, 197–202. https://doi.org/10.1038/s41586-018-0012-7.

Wickham, H., 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York.

Wood, D.E., Salzberg, S.L., 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 15, R46. https://doi.org/10.1186/gb-2014-15-3-r46.