Biochemical and Biophysical Research Communications 436 (2013) 525–529

Contents lists available at SciVerse ScienceDirect

Biochemical and Biophysical Research Communications

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

Viral categorization and discovery in human circulation by transcriptome sequencing

Weihua Wang [a,c,1], Xiaoan Zhang [a,1], Yanjuan Xu [a], Adrian M. Di Bisceglie [a,b,*], Xiaofeng Fan [a,b,*]

[a] Division of Gastroenterology and Hepatology, Department of Internal Medicine, Saint Louis University School of Medicine, St. Louis, MO 63104, USA
[b] Saint Louis University Liver Center, Saint Louis University School of Medicine, St. Louis, MO 63104, USA
[c] Wuhan Center for Tuberculosis Control, Wuhan 430030, Hubei, China

ARTICLE INFO

Article history:
Received 17 May 2013
Available online 11 June 2013

Keywords:
Viral discovery
Next-generation sequencing
Multiple displacement amplification

ABSTRACT

Serum is the most common and easily accessible patient specimen in a minimally invasive manner. As a biological resource, RNA in serum has been less explored for its clinical utilization due to prevailing concerns regarding its high degradable nature. In the current study, however, we have documented the use of human serum RNA for viral categorization and discovery through transcriptome sequencing and analysis using well-curated databases and advanced bioinformatic tools. Such an integrated approach may have an immediate application in any clinical situations concerning with viral etiology.

1. Introduction

In many clinical situations, viral infection is frequently suspected as an etiological factor. Yet unless there is an explicit viral candidate, the confirmation of putative viral infections is a difficult task. Historically viral categorization and discovery is essentially a technology-driven process. When candidate viruses are not readily grown *in vitro*, the detection of virus-encoded products or viral genomes becomes the only choice. In this setting, many methods have been developed with a focus on the high throughput nature, such as immune-based library screening [1], mass spectrometry [2], microarray [3] and next-generation sequencing (NGS) [4–6]. Among them NGS represents the most attractive approach due to its large dynamic range for gene detection and the independence of any viral sequence information [7,8]. Indeed, by taking advantage of complete decipherment of human genome sequences, a NGS-based approach, named transcriptome subtraction, had been developed and achieved initial success [4]. However, most studies, if not all, use human tissues as a starting material. In practice, tissue is not readily accessible or feasible in situations where there is no explicit target for a suspicious viral infection. Similarly, in a "hit and run" infection mode [9], there is a very narrow time window for tissue sampling. In the current study, by the integration of an enhanced amplification technique and advanced bioinformatic tools, we present a robust, sensitive and simplified NGS-based method that uses human serum as a biological source for viral categorization and discovery.

2. Materials and methods

2.1. Serum samples

In the current study, hepatitis C virus (HCV), one of the medically important RNA viruses with a single stranded RNA genome approximately at 9600 base pairs [2], was used as a model viral agent for both optimization and validation of experimental protocols.

Serum sample #1709, from a patient with chronic HCV infection, was available at large volume that allowed extensive experimental optimization. Additional serum samples, either HCV-negative or positive, were collected from patients at the Saint Louis University Hospital liver clinic. Informed consent and institutional review board approval were obtained prior to the study. All samples were stored at −80 °C until use.

2.2. Measurement of serum RNA concentration

Total RNA was extracted from 140 μL serum and eluted into 60 μL Tris buffer (pH 8.5) using QIAamp Viral RNA Mini kit (Qiagen). RNA concentration was measured with Qubit RNA BR Assay Kit in the Qubit 2.0 Fluorometer (Life Technologies). Measurement for each RNA sample was repeated three times and the mean values were used to calculate total RNA concentration in corresponding serum samples.

* Corresponding authors. Address: Division of Gastroenterology and Hepatology, Department of Internal Medicine, Saint Louis University School of Medicine, 3635 Vista Avenue, St. Louis, MO 63110, USA. Fax: +1 314 577 8125.
E-mail addresses: dibiscam@slu.edu (A.M. Di Bisceglie), fanx@slu.edu (X. Fan).
[1] These authors contributed equally to this work.

0006-291X/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.bbrc.2013.05.139

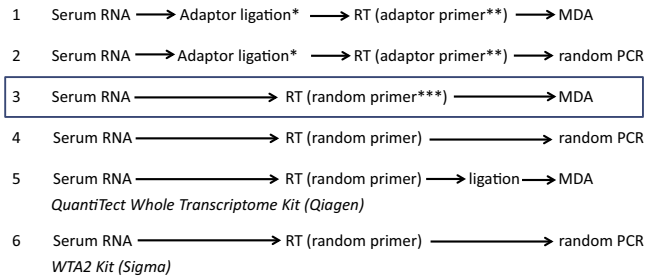

Fig. 1. A brief summary of amplification strategies. Efficient amplification of total serum RNA was estimated by six approaches, including two commercial kits (#5 and #6). The final product from each protocol was examined for robust PCR detection of HCV 5′UTR region. Only protocol #3 (framed) met this standard and was subjected to next-step optimization. *RNA ligation used a highly efficient adaptor recently developed in our lab [10]; **Adaptor primer represents the 5′ part of the adaptor [10]; ***Random primer is an exonuclease-resistant hexamer (Fidelity Systems). RT, reverse transcription; MDA, multiple displacement amplification.

2.3. Unbiased cDNA amplification from serum samples

Due to a low concentration of serum RNA, an amplification step after RT is necessary prior to NGS. There are currently no existing protocols that demonstrate an unbiased amplification from extracted serum RNA, an extremely heterogeneous sample type. In the current study, such an unbiased amplification was achieved through a two-step optimization strategy, the determination of the best approach and a further optimization of the defined approach.

2.4. Approaches for unbiased serum cDNA amplification

A total of six experimental approaches, including two commercial kits, were empirically decided to estimate their ability for an unbiased amplification of serum cDNA (Fig. 1). Approaches #1 and #2 had a ligation step prior to RT, which was accomplished with a powerful adaptor Linker 2 as we described previously [10]. An aliquot of 5 μL ligation product was then mixed with 15 μL RT matrix to formulate RT reaction containing 1× Mg^{2+}-free SuperScript III buffer, 5 mM DTT, 1 mM dNTPs (New England Biolabs), 16 U of Rnasein (Promega), 1 mM reverse primer HBVR1linker2 [10] and 200 U SuperScript III (Life Technologies), followed by 1-h incubation at 50 °C.

The RT product was used either for multiple displacement amplification (MDA) (approach #1) or random PCR (approach #2) (Fig. 1). The MDA was conducted in a 50-μL reaction volume consisting of 5 μL RT product, 1× phi29 DNA polymerase reaction buffer, 1 mM dNTPs, 30 U of phi29 DNA polymerase (New England Biolabs). The reaction was incubated at 30 °C for 12 h.

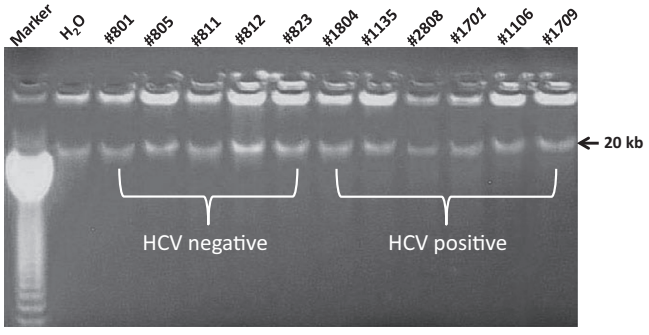

Fig. 2. A representative run of RT/MDA product from sample #1709 and additional 10 samples with or without HCV infection. A dominant band at approximate 20 kb was consistently observed for each sample as well as the negative control (H_2O).

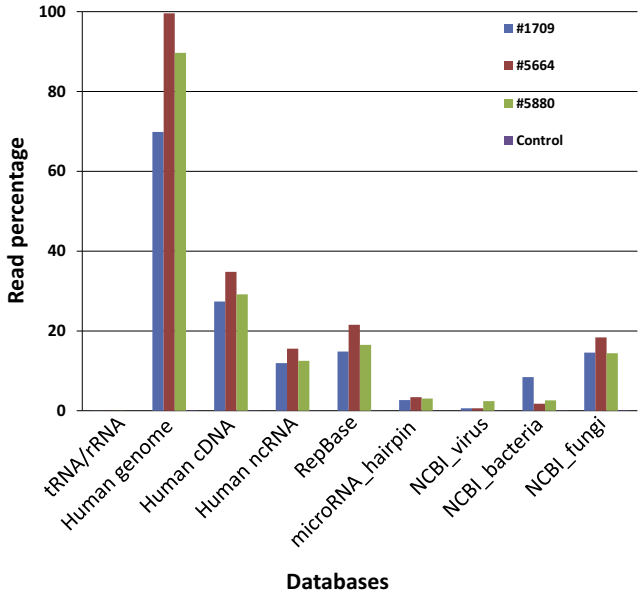

Fig. 3. Read mapping against databases by Bowtie 2 under default mapping score setting. Reads from each sample were first processed with quality control and duplicate removal. All three samples showed generally comparable mapping rates to individual databases. The negative control (H_2O) had almost negligible mapping against databases. Unlike cell-based transcriptome data, tRNA/rRNA was almost undetected. Mature microRNAs were not detected due to the size exclusion (<50 bp) in serum RNA extraction by QIAamp Viral RNA Mini kit.

In random PCR approach, entire RT product was mixed with 30-μL PCR matrix containing 3 μL of 10× DyNAzyme buffer, 0.4 μL of forward primer random 656 (5′-tac agc cta ctc cca tct ctc cac cn_7t ggc-3′) and 1 U DyNAzyme EXT DNA polymerase (ThermoFisher Scientific). Cycle parameters on DNA 480 cycler were adapted from our long RT-PCR technique except for the extension time reduced to 1 min [11]. An aliquot of 2 μL 1st round product was used for the 2nd PCR with primers 3HBVF1 and 3HBVR1 [11].

Approaches #3 and #4 were respectively similar to #1 and #2 except for the omission of RNA ligation step (Fig. 1). Two commercial kits, QuantiTect Whole Transcriptome Kit (Qiagen) and WTA2 Kit (Sigma), were also included for the estimation of unbiased cDNA amplification. The major experimental steps were outlined (Fig. 1) and their performance was basically according to instructions from the manufacturers.

The approach with robust detection of HCV 5′ non-translated domain (5′UTR) from repeated experiments was progressed into next-step optimization. Based on this selection criteria, the RT/MDA protocol (approach #3) was subjected to further optimization for an unbiased amplification of all components in serum RNA samples, as indicated by simultaneous detection in their final products for four HCV genes over entire HCV genome, i.e., 5′ NTR, Core, NS3 and NS5a (Table S1). Experimental optimization focused on the adjustment of three major parameters, including time length (from 4 to 22 h) of MDA incubation, primer concentrations and reaction buffers.

2.5. Pyrosequencing

Using the optimized RT/MDA protocol, 10 μL of extracted serum RNA was used as starting material for cDNA amplification in sample #1709 and two additional HCV-negative samples, #5664 and #5880. A negative control (water) was also included. The RT/MDA product was purified using QIAamp DNA Mini Kit (Qiagen)

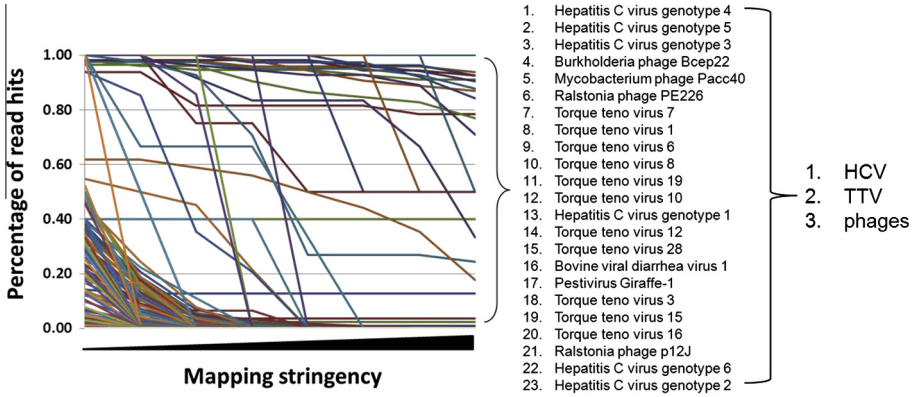

Fig. 4. Virus categorization in sample #1709. After the quality control, reads were directly mapped against NCBI virus reference sequences using a gapped aligner Bowtie 2. An enrichment strategy was applied by modulating mapping stringency through functional mapping score setting. The number of read hits under the default (lowest) mapping stringency was used as the base to calculate declining rates of each virus-mapped read. Most mapped reads were dropped with gradually enhanced mapping stringency. With a cutoff >0 at the highest mapping stringency, such an enrichment strategy returned 23 viruses. By taking into account viral genetic diversity, sample #1709 was circulated with HCV, Torque teno virus (TTV) and phages. The enrichment strategy thus eliminated false viral detection and provided reliable viral categorization resolved at a single read level.

and quantified by Qubit Fluorometer (Invitrogen). An aliquot of 5 μg RT/MDA product from each sample and the control was used for shotgun library construction and then subjected to pyrosequencing on Roche 454 GS/FLX Titanium platform.

2.6. Sequence analysis

2.6.1. Data quality control

Raw sequencing reads were filtered in program PRINSEQ (v0.19.5) for quality control, including read length ($\leqslant$50 bp), mean read quality score ($\leqslant$25), low complexity with DUST score $\geqslant$7 and ambiguous bases ($\geqslant$1%) [12]. PRINSEQ was also used for basic read statistics such as the distribution of read length, quality score, GC content and overall data pattern [12]. Filtered reads were examined for their compositions and relative abundance by mapping using a gapped aligner Bowtie 2 [13] against various databases, including human genome UCSC hg19 [14], human cDNA and non-coding RNA (Ensemble release 69) [15], microRNA (mature and hairpin) (release 19) [16], human repetitive DNA elements (RepBase, v17.11) [17], NCBI reference sequences (bacteria, 7.6 GB; virus, 94.4 MB, fungi 3.2 GB) [18] and an in-house human tRNA/rRNA collection (3771 sequences) derived from UCSC human genome and SLIVA [19].

2.6.2. Viral categorization

Categorization of known viruses utilized NCBI virus reference database consisting of 4352 entire viral genomes in its most recent release. An enrichment strategy was applied to minimize false positive detection and to increase the sensitivity. Briefly, mapping stringency was gradually increased in Bowtie 2 through a functional score setting $f(x) = n + 8\ln(x)$ where "x" indicated the read length and "n" was adjusted from 20 (default setting) to 120 by a "10" interval. Read hits from individual virus genome were counted at each score setting in SAM output through either SAMtools (idxstats) [20] or BEDTools [21]. Declining rates of read hits for each mapped viral genome were then calculated against the baseline scoring setting (default setting), which generated a map of the declining of all mapped read hits for viral categorization. Under the highest mapping score setting, i.e., [$f(x) = 120 + 8\ln(x)$], a long read at 800 bp requires score at least 173.4 for an effective mapping that is equal to 87-bp complete match or extended lengths with penalty from mismatches [13]. Assuming a cutoff value >0 at the highest score setting, the detection of a given virus by our enrichment strategy thus depends on the availability of at least one read hit in mapping at the highest score setting.

2.6.3. Approach to viral discovery

For discovery of possible novel viruses, unmapped reads were obtained through successive subtractive mapping in Bowtie 2 against the read library from the control and databases from human and microbes. Reads with >95% sequence identity were further removed by cd-hit-454 [22]. Resulting reads were examined for hits in NCBI "nr" protein sequence database by BLASTX with an empirical E value setting at 0.001 [23]. Next, all open reading frames (ORFs) longer than 50 amino acids in six frames were extracted from remaining reads by EMBOSS getorf module [24], followed by the scanning of protein signatures using InterProScan (v5RC4) [25] against the InterPro protein family database [26]. All resulting ORFs were analyzed by BLASTP with E value at 0.01 against NCBI viral protein database generated by MAKEBLASTDB program. PSI-BLAST was used as the last step to find evolutionally distant viral relatives in final set of ORFs [23]. Alternatively, unmapped reads were directly used for ORF extraction, followed by BLASTP and PSI-BLAST analyses against viral protein database (Fig. S1).

2.6.4. Data availability

Raw sequence data from three samples and the control were archived in NCBI Sequence Read Archive (SRA) under SRA accession number SRA065358.

3. Results

3.1. Unbiased amplification of serum total RNA molecules

Assuming complete RNA recovery, the quantitation of extracted RNA from 30 serum samples was translated into an average serum RNA concentration at 0.71 ± 0.17 ng/μL, ranging from 0.34 to 1.14 ng/μL. Sample #1709 had a total serum RNA concentration at 0.64 ng/μL and a typical HCV RNA titer at 3.0×10^6 copies/mL as determined by Roche Amplicor HCV Monitor (version 2.0). Thus HCV RNA in this patient accounted for approximately 1/10,000th of total serum RNA amount. Such a low concentration, together with its long genome size (~9600 bp) in comparison to most human cDNA transcripts, made HCV genome to be a natural spike RNA to monitor the efficiency of unbiased serum cDNA amplification.

Among six approaches tested for sequence-independent amplification, only approach #3 achieved the robust detection of HCV 5′ UTR in repeated experiments (Fig. 1). Subsequent optimization based on this approach formulated a protocol that allowed simultaneous detection of four separate regions of the HCV genome from RT/MDA product as well as the combination of RT and MDA into a single tube. Briefly, 10 μL of extracted serum RNA was mixed with 10 μL of RT matrix in 20 μL reaction containing 4 mM $MgCl_2$, 5 mM DTT, 1 μM of exonuclease-resistant hexamer, 2 mM dNTPs, 200 units of SuperScript III reverse transcriptase, 50 mM Tris–HCl and 15 mM $(NH4)_2SO_4$ with a final buffer pH at 8.5. After incubation at 50 °C for 60 min, 4 μL of 20 μM exonuclease-resistant hexamer was added into the tube and denatured at 94 °C for 3 min, followed by the addition of 26 μL MDA matrix consisting of 2.4 μL of 50 mM $MgCl_2$, 3 μL of 10× phi29 buffer, 0.25 μL of 40 mM dNTPs and 30 U of phi29 DNA polymerase. The reaction was incubated at 30 °C for 18 h and then heat inactivated.

The simplified RT/MDA protocol gave a robust amplification of all serum samples with yields in microgram quantities (Fig. 2). Four HCV regions were detected in the RT/MDA product from all HCV-positive serum samples (data not shown). The strong strand-displacement activity of phi29 DNA polymerase made final products to form a hyper-branched structure dominant at about 20 kb (Fig. 2). Such a product was even observed in the blank control (water) (Fig. 2), representing potential artifacts from the polymerization of hexamer dimers by phi29 DNA polymerase, which was reported to bind and initiate polymerization of as little as 4 bp of oligonucleotide in spite of a less stable status [27]. Indeed, the product from the negative control was completely eliminated when setting the temperature of MDA at 34 °C (data not shown). However, raising the temperature for MDA above 30 °C resulted in a dramatic decrease of the yield of final product.

3.2. Data overview

Sample #1709 had an average read length at 413.9 bp that was longer than the other libraries (Table S2), perhaps due to batch-to-batch differences in library preparation and pyrosequencing. After the read quality control and the removal of primer artifacts, all three samples showed similar GC content and data patterns (data not shown). It is worthwhile to note a high percentage of primer artifacts in sample #1709 (72.6%) comparing to samples #5664 (7.17%) and #5880 (12.55%) (Table S2). As sample #1709 had the lowest serum RNA concentration among three samples (Table S2), this observation confirmed a previous report that low concentration of templates in MDA stimulated the generation of primer artifacts [28]. After the exclusion of primer artifacts, read duplicates were higher in #1709 than samples #5664 and #5880, a common finding in metagenomic data believed to be associated with sequence depths [22].

The negative control had almost negligible mapping rates against all databases (Fig. 3). As expected, individual database mapping revealed the dominance of reads as being of human origin in each patient sample, however, a small but significant part of the reads could be mapped to microbes (Fig. 3). Yet accumulating mapping rates of human and microbes exceeded 100% for each sample data even under the higher mapping stringency, suggesting multi-hit characteristics of at least a subset of reads in each dataset. In this setting, chimeric reads, containing sequences from different species, could be an explanation due to the use of phi29 DNA polymerase in MDA. Its high strand-displacement activity may facilitate the generation of chimeric product from genetically heterogeneous templates [29].

3.3. Categorization of known viruses

Viral categorization was done by direct read mapping against NCBI virus database using gapped aligner Bowtie 2 [13]. Because of a potential occupation of chimeric or multi-hit reads in datasets, Bowtie 2 was run under "–a" mode that reported all alignments under defined score settings. To eliminate false positive detection resulting from the sequence similarity between humans and microbes [30], an enrichment strategy was applied by modulating the mapping stringencies. In sample #1709, enrichment with a cutoff >0 at the highest score setting returned 23 viruses, including six major HCV genotypes, eleven Torque teno virus (TTV) subtypes and four phages (Fig. 4). TTV was also a major virus in circulation in samples #5664 and #5880, as revealed by similar enrichment analysis (Fig. S2). It should be noted that sequence similarity among viruses should be taken into account in final data interpretation. For a given virus, NCBI virus reference database includes its major serotypes or genotypes. These serotypes or genotypes may be all recruited, as seen for HCV and TTV in case of sample #1709 (Fig. 4). Thus the current categorization strategy cannot resolve viral isolates at serotype or genotype level. Further determination of viral serotypes or genotypes requires phylogenetic analysis of individual reads.

3.4. Discovery of novel viruses

Using step-wise pipelines described for viral discovery, sample #1709 had 14,474 reads remained after subtractive mapping, the removal of duplicates and length exclusion (>100 bp). Among these 14,474 reads, 10,846 (74.9%) reads had at least one hit by BLASTX in all six frames against NCBI "nr" non-redundant protein database. By combination, these reads actually targeted 2793 proteins, mostly from bacteria (Table S3). The only virus-related hits were from TTV (Table S3).

For remaining 3628 reads, a total of 14,540 ORFs longer than 50 amino acids were extracted in all six frames using EMBOSS getorf script [24]. These ORFs were scanned for protein signatures by InterProScan and only 40 ORFs from 39 reads had a positive result, suggesting that most of 3628 reads may represent sequencing or primer artifacts. By BLASTP analysis against the NCBI nr database and viral protein database with respective E values at 0.001 and 0.01, two ORFs had bacteria-relevant hits and 13 ORFs were associated with phages. The remaining 25 ORFs were used for PSI-BLAST analysis with viral protein database. With the E value setting at 10, all of 25 ORFs got protein hits mostly associated with phages. One ORF with corresponding read identifier GG6NFNQ01C9RIF was found to be relevant to human T-lymphotropic virus and thus warranted further investigation.

4. Discussion

The current study systematically documented an integrated platform for viral categorization and discovery from human circulation. First, using HCV as a model agent, unbiased amplification of minute amounts of serum RNA was achieved by optimized RT/MDA, supporting a general observation that MDA outperforms PCR-based amplification in terms of linear characteristics [31]. The combination of RT with MDA into a single tube further simplified experimental procedures, which minimizes contamination potential and enhances performance robustness (Figs. 2 and 3).

Second, the platform takes advantage of the most recent bioinformatic advances in dealing with NGS data. Due to genetic heterogeneity of serum RNA, both MDA-based amplification and NGS itself may result in the formation of chimeric reads, as indicated by our mapping results (Fig. 3) [29]. To prevent the loss of

potentially valuable information from chimeric reads, all alignments found by Bowtie 2 were exported for downstream enrichment, which not only eliminated false positive detection due to sequence similarity but also maximized the detection sensitivity. In this setting, it should be noted there are two special cases in which viruses have exceptional sequence similarity with the human genome, including bovine viral diarrhea virus 1/pestivirus Giraffe-1 and murine osteosarcoma virus due to an insertion of cellular sequences and the existence of cellular homolog, respectively [32,33]. These viruses should be removed from the list of viral categorization as seen in sample #1709 (Fig. 4). Under our enrichment strategy, categorization of any known viruses could be resolved at a single read level. Similar resolution was also achieved for viral discovery using either of pipelines (Fig. S1). Importantly, direct BLASTP analysis using extracted ORFs from unmapped reads represents a more computationally efficient approach due to the omission of BLASTX, a bottleneck in processing of large NGS data. In a re-analysis of two read libraries that resulted in the discovery of Merkel Cell polyomavirus (MCV) [4], direct BLASTP search of extracted ORFs identified three reads with statistically significant hits. Of these, two reads encoded a protein product corresponding to large T antigen of polymavirus (Table S4). Giving intrinsic heuristic algorithms [34], BLASTP analysis against a small database is thus a rapid and more sensitive method to identify virus-relevant reads.

In spite of a high resolution for either viral categorization or discovery at single real level, the performance of current platform could be further improved by several options, including the use of latest Illumina's NGS platform for deeper sequence coverage in a cost-effective manner, the increase of input RNA amount at RT/MDA to suppress the formation of primer artifacts and the application of technical replicates, in particular under a shallow sequencing depth, as RNA-Seq is essentially a sampling procedure [35].

In summary, human serum is easily accessible in a minimally invasive manner and appears to show considerably stability at appropriate storage conditions [36,37]. Using serum RNA as starting material, the present study has documented a technical platform for efficient categorization of both RNA (e.g. HCV) and DNA viruses (e.g. TTV) and viral discovery resolved at single-read level. The robust single-tube amplification and improved bioinformatic pipelines allow this platform to be of immediate application in clinical situations concerning an unknown viral infection.

Acknowledgments

We thank Carie A. Tebbe (Saint Louis University) for assistance with the Red Hat Linux operational system. We also thank Drs. Robert Schmieder (San Diego State University), Aaron R. Quinlan (University of Virginia School of Medicine) and John Urban (Brown University) for helpful discussions on data analysis. This work was supported by NIH grants R01 DK80711 (XF) and R21 AI076834 (AMD).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.bbrc.2013.05.139.

References

[1] Q.L. Choo, G. Kuo, A.J. Weiner, L.R. Overby, D.W. Bradley, et al., Isolation of a cDNA clone derived from a blood-borne non-A, non-B viral hepatitis genome, Science 244 (1989) 359–362.
[2] Y. Ye, E.C. Mar, S. Tong, S. Sammons, S. Fang, et al., Application of proteomics methods for pathogen discovery, J. Virol. Methods 163 (2010) 87–95.
[3] D. Wang, A. Urisman, Y.T. Liu, M. Springer, T.G. Ksiazek, et al., Viral discovery and sequence recovery using DNA microarrays, PLoS Biol. 1 (2003) E2.
[4] H. Feng, M. Shuda, Y. Chang, P.S. Moore, Clonal integration of a polyomavirus in human merkel cell carcinoma, Science 319 (2008) 1096–1100.
[5] G. Palacios, J. Druce, L. Du, T. Tran, C. Birch, et al., A new arenavirus in a cluster of fatal transplant-associated diseases, N. Engl. J. Med. 358 (2008) 991–998.
[6] T. Briese, J.T. Paweska, L.K. McMullan, S.K. Hutchison, C. Street, et al., Genetic detection and characterization of Lujo virus, a new hemorrhagic fever associated arenavirus from Southern Africa, PLoS Pathog. 5 (2009) e1000455.
[7] R.A. Moore, R.L. Warren, J.D. Freeman, J.A. Gustavsen, C. Chénard, et al., The sensitivity of massively parallel sequencing for detecting candidate infectious agents associated with human tissue, PLoS One 6 (2011) e19838.
[8] P. Tang, C. Chiu, Metagenomics for the discovery of novel human viruses, Future Microbiol. 5 (2010) 177–189.
[9] G.R. Skinner, Transformation of primary hamster embryo fibroblasts by type 2 simplex virus: evidence for a "hit and run" mechanism, Br. J. Exp. Pathol. 57 (1976) 361–376.
[10] X. Zhang, X. Fan, Y. Xu, A.M. Di Bisceglie, Enhanced protocol for determining the 3′ terminus of hepatitis C virus, J. Virol. Methods 167 (2010) 158–164.
[11] X. Fan, Y. Xu, A.M. Di Bisceglie, Efficient amplification and cloning of near full-length hepatitis C virus genome from clinical samples, Biochem. Biophys. Res. Commun. 346 (2006) 1163–1172.
[12] R. Schmieder, R. Edwards, Quality control and preprocessing of metagenomic datasets, Bioinformatics 27 (2011) 863–864.
[13] B. Langmead, S.L. Salzberg, Fast gapped-read alignment with bowtie 2, Nat. Methods 9 (2012) 357–359.
[14] W.J. Kent, C.W. Sugnet, T.S. Furey, K.M. Roskin, T.H. Pringle, et al., The human genome browser at UCSC, Genome Res. 12 (2002) 996–1006.
[15] P. Flicek, M.R. Amode, D. Barrell, K. Beal, S. Brent, et al., Ensembl 2012, Nucleic Acids Res. 40 (2012) D84–D90.
[16] A. Kozomara, S. Griffiths-Jones, MiRBase: integrating microRNA annotation and deep-sequencing data, Nucleic Acids Res. 39 (2011) D152–D157.
[17] J. Jurka, V.V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany, et al., Repbase update, a database of eukaryotic repetitive elements, Cytogenet. Genome Res. 110 (2005) 462–467.
[18] K.D. Pruitt, T. Tatusova, G.R. Brown, D.R. Maglott, NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy, Nucleic Acids Res. 40 (2012) D130–D135.
[19] C. Quast, E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, et al., The SILVA ribosomal RNA gene database project: improved data processing and web-based tools, Nucleic Acids Res. 41 (2013) D590–D596.
[20] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, et al., The sequence alignment/map format and SAMtools, Bioinformatics 25 (2009) 2078–2079.
[21] A.R. Quinlan, I.M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features, Bioinformatics 26 (2010) 841–842.
[22] B. Niu, L. Fu, S. Sun, W. Li, Artificial and natural duplicates in pyrosequencing reads of metagenomic data, BMC Bioinformatics 11 (2010) 187.
[23] S.F. Altschul, T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, et al., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389–3402.
[24] P. Rice, I. Longden, A. Bleasby, EMBOSS: the European molecular biology open software suite, Trends Genet. 16 (2000) 276–277.
[25] E. Quevillon, V. Silventoinen, S. Pillai, N. Harte, N. Mulder, et al., InterProScan: protein domains identifier, Nucleic Acids Res. 33 (2005) W116–W120.
[26] S. Hunter, P. Jones, A. Mitchell, R. Apweiler, T.K. Attwood, InterPro in 2011: new developments in the family and domain prediction database, Nucleic Acids Res. 40 (2012) D306–D312.
[27] A.J. Berman, S. Kamtekar, J.L. Goodman, J.M. Lázaro, M. de Vega, et al., Structures of phi29 DNA polymerase complexed with substrate: the mechanism of translocation in B-family polymerases, EMBO J. 26 (2007) 3494–3505.
[28] R.S. Lasken, Genomic DNA amplification by the multiple displacement amplification (MDA) method, Biochem. Soc. Trans. 37 (2009) 450–453.
[29] R.S. Lasken, T.B. Stockwell, Mechanism of chimera formation during the multiple displacement amplification reaction, BMC Biotechnol. 7 (2007) 19.
[30] G. Fan, J. Li, Regions identity between the genome of vertebrates and non-retroviral families of insect viruses, Virol. J. 8 (2011) 511.
[31] J. Kim, C.J. Easley, Isothermal DNA amplification in bioanalysis: strategies and applications, Bioanalysis 3 (2011) 227–239.
[32] P. Becher, H.J. Thiel, M. Collins, J. Brownlie, M. Orlich, Cellular sequences in pestivirus genomes encoding gamma-aminobutyric acid (A) receptor-associated protein and Golgi-associated ATPase enhancer of 16 kilodaltons, J. Virol. 76 (2002) 13069–13076.
[33] T. Curran, W.P. MacConnell, F. van Straaten, I.M. Verma, Structure of the FBJ murine osteosarcoma virus genome: molecular cloning of its associated helper virus and the cellular homolog of the v-fos gene from mouse and human cells, Mol. Cell Biol. 3 (1983) 914–921.
[34] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410.
[35] L.M. McIntyre, K.K. Lopiano, A.M. Morse, V. Amin, A.L. Oberg, et al., RNA-seq: technical variability and sampling, BMC Genomics 12 (2011) 293.
[36] N.B. Tsui, E.K. Ng, Y.M. Lo, Stability of endogenous and added RNA in blood specimens, serum, and plasma, Clin. Chem. 48 (2002) 1647–1653.
[37] P.B. Gahan, M. Stroun, The biology of circulating nucleic acids in plasma and serum (CNAPS), Nucleic Acids Mol. Biol. 25 (2010) 167–189.