Contaminating viral sequences in high-throughput sequencing viromics: a linkage study of 700 sequencing libraries

Asplund, Maria; Kjartansdóttir, Kristín Rós; Mollerup, Sarah; Vinner, Lasse; Fridholm, Helena; Herrera, José A. R.; Friis-Nielsen, Jens; Hansen, Thomas Arn; Jensen, Randi Holm; Nielsen, Ida Broman

Total number of authors:
23

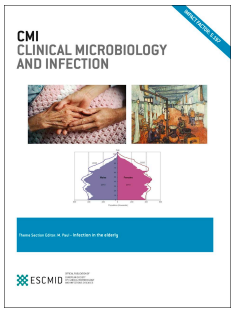
Published in:
Clinical Microbiology and Infection

Link to article, DOI:
10.1016/j.cmi.2019.04.028

Publication date:
2019

Document Version
Peer reviewed version

Accepted Manuscript



Contaminating viral sequences in high-throughput sequencing viromics: a linkage study of 700 sequencing libraries

Maria Asplund, Kristín Rós Kjartansdóttir, Sarah Mollerup, Lasse Vinner, Helena Fridholm, José A.R. Herrera, Jens Friis-Nielsen, Thomas Arn Hansen, Randi Holm Jensen, Ida Broman Nielsen, Stine Raith Richter, Alba Rey-Iglesia, Maria Luisa Matey-Hernandez, David E. Alquezar-Planas, Pernille.V.S. Olsen, Thomas Sicheritz-Pontén, Eske Willerslev, Ole Lund, Søren Brunak, Tobias Mourier, Lars Peter Nielsen, Jose.M.G. Izarzugaza, Anders Johannes Hansen

PII: S1198-743X(19)30206-X

DOI: https://doi.org/10.1016/j.cmi.2019.04.028

Reference: CMI 1656

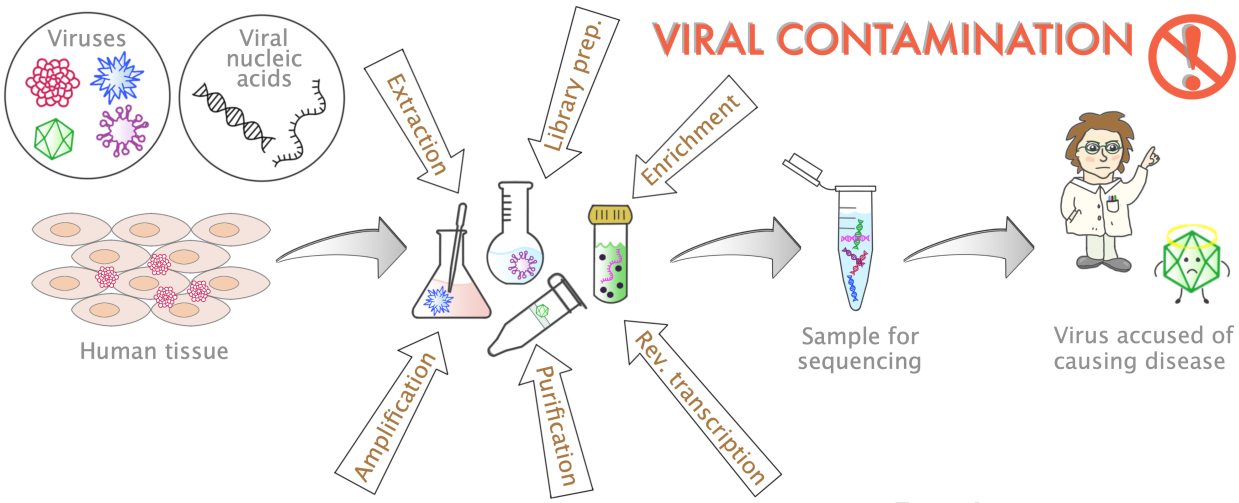
To appear in: *Clinical Microbiology and Infection*

Received Date: 3 January 2019

Revised Date: 12 April 2019

Accepted Date: 18 April 2019

Please cite this article as: Asplund M, Kjartansdóttir KR, Mollerup S, Vinner L, Fridholm H, Herrera JAR, Friis-Nielsen J, Hansen TA, Jensen RH, Nielsen IB, Richter SR, Rey-Iglesia A, Matey-Hernandez ML, Alquezar-Planas DE, Olsen PVS, Sicheritz-Pontén T, Willerslev E, Lund O, Brunak S, Mourier T, Nielsen LP, Izarzugaza JMG, Hansen AJ, Contaminating viral sequences in high-throughput sequencing viromics: a linkage study of 700 sequencing libraries, *Clinical Microbiology and Infection*, https://doi.org/10.1016/j.cmi.2019.04.028.



Contaminating viral sequences in high-throughput sequencing viromics: a linkage study of 700 sequencing libraries

Maria Asplund[1*], Kristín Rós Kjartansdóttir[1], Sarah Mollerup[1], Lasse Vinner[1], Helena Fridholm[1], José A. R. Herrera[3,4], Jens Friis-Nielsen[4], Thomas Arn Hansen[1], Randi Holm Jensen[1], Ida Broman Nielsen[1], Stine Raith Richter[1], Alba Rey-Iglesia[1], Maria Luisa Matey-Hernandez[4], David E. Alquezar-Planas[1], Pernille V. S. Olsen[1], Thomas Sicheritz-Pontén[1,5], Eske Willerslev[1], Ole Lund[4], Søren Brunak[3,4], Tobias Mourier[1], Lars Peter Nielsen[2], Jose M. G. Izarzugaza[4] and Anders Johannes Hansen[1*]

[1] Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, DK-1350 Copenhagen, Denmark.

[2] Department of Autoimmunology and Biomarkers, Statens Serum Institut, DK-2300 Copenhagen S, Denmark.

[3] Disease Systems Biology Program. Panum Instituttet, Blegdamsvej 3, DK-2200 Copenhagen, Denmark

[4] Department of Bio and Health Informatics, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark.

[5] Centre of Excellence for Omics-Driven Computational Biodiscovery, AIMST University, Kedah, Malaysia

* Correspondence:

amasplund@snm.ku.dk; Tel.: +45 26 97 72 94

ajhansen@snm.ku.dk; Tel.: +45 28 75 61 34

Keywords: Contamination, laboratory component, next generation sequencing, high-throughput sequencing, virus, virome, nucleic acid, metagenomic, cluster.

Abstract

Objectives

Sample preparation for High-throughput sequencing (HTS) includes treatment with various laboratory components, potentially carrying viral nucleic acids, the extent of which has not been thoroughly investigated. Our aim was to systematically examine a diverse repertoire of laboratory components used to prepare samples for HTS in order to identify contaminating viral sequences.

Methods

A total of 322 samples of mainly human origin were analysed using eight protocols, applying a wide variety of laboratory components. Several samples (60% of human specimens) were processed by different protocols. In total 712 sequencing libraries were investigated for viral sequence contamination.

Results

Among sequences showing similarity to viruses, 493 were significantly associated to the use of laboratory components. Each of these viral sequences showed sporadic appearance, only being identified in a subset of the samples treated with the linked laboratory component, and some were not identified in the non-template control (NTC) samples. Remarkably, more than 65% of all viral sequences identified were within viral clusters linked to the use of laboratory components.

Conclusions

We show that high prevalence of contaminating viral sequences can be expected in HTS-based virome data and provide an extensive list of novel contaminating viral sequences that can be used for evaluation of viral findings in future virome and metagenome studies. Moreover we show that detection can be problematic due to stochastic appearance and limited NTCs. Although the exact origin of these viral sequences requires further research, our results support laboratory component linked viral sequence contamination of both biological and synthetic origin.

Introduction

High-throughput sequencing (HTS) is an indispensable tool in life science research and clinical diagnostics (1,2) and facilitates the generation of massive amounts of DNA sequence information at acceptable costs within a short timeframe. The field of viromics has benefited from the rapid improvement of HTS technologies, as evidenced by major discoveries of novel viruses (3-9), some of which have proven to be the cause of recent human epidemics (10). Due to high sequence diversity, it has been challenging to identify novel viral genomes in clinical specimens using sequence-specific molecular methods such as PCR. HTS technologies provide an attractive alternative approach for virus discovery that require no prior knowledge about viral genomes. However, discovery of viruses using HTS also poses a number of challenges that must be accounted for in the interpretation of data. Sample preparation for HTS includes treatment with various laboratory components, also used for sample preparation in other non-HTS methods. Laboratory components have previously been documented to carry viral nucleic acid contamination (11-17). Great caution is therefore essential when claiming disease association with a particular microorganism, to avoid incorrect conclusions, as in some unfortunate recent examples (18-22). A better understanding of laboratory component derived contamination is needed.

Here, we systematically address the problem of nucleic acid contamination using HTS with focus on virus identification in clinical samples. We provide a comprehensive in silico characterization of contaminating viral sequences and their probable sources. More than 300 samples were analysed using eight overall methods applying an extensive variety of laboratory components. The original purpose of the investigation was to identify sample-derived viral sequences, findings described in Mollerup et al. (submitted elsewhere). In many cases (165 out of 274 human specimens), the same sample was processed by different laboratory protocols resulting in several sequencing libraries per sample. Consequently, this study poses a unique opportunity for the characterization of common viral artefacts and contaminants in HTS metagenomic studies within clinical and other samples.

Methods

Ethics statement

The Regional Committee on Health Research Ethics and the National Committee on Health Research Ethics decided that ethical permission was not needed for collection and processing of these samples (case no. H-2-2012-FSP2 and 1304226) according to the Danish national legislation (Sundhedsloven). The samples used in this study were processed anonymously. All experiments were conducted according to the Declaration of Helsinki.

Samples

Samples consisted of 274 human specimens (32 different sample types, mostly of cancerous origin), 5 virus-spiked positive control samples and 43 non-template controls (see Supplementary material, Table S1). Laboratory method development was not part of this study and positive controls were included to assess bioinformatic pipeline.

Sample processing

In order to identify viral sequences within the samples, eight different overall methods were used; four DNA and four RNA focused methods (shotgun DNA and RNA, circular DNA enrichment, virion enrichment DNA and RNA, mRNA enrichment, retrovirus capture DNA and mRNA) (see Supplementary material, Laboratory methods S1). A total of 712 sequencing libraries were prepared and sequenced on the Illumina Hiseq 2000 platform with 2x100bp paired-end sequencing. For sample processing 54 laboratory reagents and utilities (laboratory components) were applied (see Supplementary material, Fig. S1 and Table S2). All samples were processed in the same laboratory.

Characterization of sequencing data

Paired-end sequencing reads were adapter trimmed and quality trimmed and merged. Reads mapping to the human reference genome (hg38), reads of length <30 nucleotides, and low-complexity reads were excluded from further analysis. Remaining reads were assembled into

larger contiguous sequences (contigs) from a combination of pairs, collapsed (merged overlapping pairs) and singleton reads. Default parameters were used for this purpose. Contigs and all human depleted and quality filtered reads were queried against the NCBI nucleotide database (nt) using BLASTn (megablast) (23) with a cut-off e-value of 10^{-3}. Contigs with no BLASTn hit were queried against the NCBI non-redundant protein database (nr) using BLASTx with the same cut-off e-value. For each characterized sequence the best hit was selected and taxonomically classified using the NCBI taxonomy database. All sequences with a viral classification were selected and sequences with the same viral taxID at the first level (species/strain) were clustered. Reads possibly occurring because of library misidentification, as a result of mixed sequencing clusters, referred to as bleedover (24), was considered. A bleedover ratio was calculated by dividing the viral read count of each viral sequence with the highest viral read count for the same viral sequence from different libraries sequenced on the same lane. Identified viral sequences with bleedover ratio lower than 0.3% were removed. Cross-contamination from one sequencing run to another was not considered but could potentially also be present. Hosts of viruses were recovered from the NCBI taxonomy browser. Statistical analysis and visualization of data was done using the software R v. 3.5.1 (25).

Association analysis

The identified viral sequences were correlated to laboratory components and sample types to detect possible sources of contamination. This was done using a positive one-tailed Fisher's exact test (significance level α=0.05 with Bonferroni correction).

Coverage analysis

A reference genome was selected for each viral sequence linked to a laboratory component (see Supplementary material, Table S3). Using Bowtie2 v. 2.2.5 (26) human depleted and quality filtered reads were mapped to viral reference genomes, applying global end-to-end and local alignment. Independently, the same reads were mapped to 6 manually selected algae chloroplast genomes. The alignments of reads to the reference genomes was visualized using Circos (v0.67-7) (27), and an additional analysis of correlation to features based on mapping results was conducted.

Cross-library genome coverage was investigated by summing the library specific genome coverage of all libraries.

Results

A diversity of viral enrichment methods was applied to 279 samples of mainly human cancerous origin, resulting in 712 sequencing libraries (see Table 1).

Table 1. Samples and libraries included in the study.

The table shows the number of samples for each sample type, the number of samples processed with the different laboratory methods, and the resulting number of libraries for each sample type (rightmost column) and laboratory method (bottom line). NTC: non-template control.

Sample type	Samples	Shotgun DNA	Shotgun RNA	Virion enrichment		Circular DNA enrich.	Capture		mRNA enrich.	Libraries
				DNA	RNA		Retro-virus DNA	Retro-virus mRNA		
Basal cell carcinoma (cutaneous)	11	11		11	11	4	6			43
Mycosis fungoides (cutaneous)	11	11		11	11	10	11			54
Melanoma (cutaneous)	10	10		10	10	8				38
Oral cancer	10	12		10	10	10				42
Oral healthy	1					1				1
Vulvar cancer	3			3	4	3				10
Bladder cancer	7			8	9	5				22
Bladder cancer urine	11		2			10				12
Colon cancer	32	12	11	3	3		6		6	41
Colon cancer blood	8	8								8
Colon cancer ascites	1	1					1			2
Colon healthy	2								2	2
Breast cancer (ductal)	10	10	10	9	13	8				50
Breast cancer (lobular)	10	10	9	10	10	7				46
Breast cancer ascites	2	1	1	1	1	2				6
Testicular cancer (seminoma)	11	1		11	12					24
Testicular cancer (non-seminoma)	5	3		5	8					16
Testicular cancer (seminoma and non-seminoma)	4	1		4	4					9
AML	15		6	9	9	7				31
B-CLL	17		8	9	9	8				34
BCP-ALL	8			8	8	8				24
CML	20		10	10	10	10				40
T-ALL	20		9	11	11	9				40
Ovarian cancer ascites	10	5	4	3	3	5				20
Pancreatic cancer ascites	4	2	2				1			5
Optic neuritis cerebrospinal fluid	4			4						4
Optic neuritis plasma	4			4						4
Vasculitis	4			4						4
Gynecological observation ascites	1		1							1
Cell lines	18	12						6	6	24
Positive control	5	10					2			12
NTC				20	18	5				43
Total	279	120	73	178	174	120	27	6	14	712

A total of 56,728,213,824 sequencing reads were generated. After human depletion and quality filtering 2,953,972,594 reads and 1,381,107 contigs were characterized using BLAST. The results are summarized in Table 2 (for library-specific information see Supplementary material, Table S1).

Table 2. Overview of the number of sequences analysed by BLAST.

	Reads	Contigs
Sequences analysed by BLAST	2,953,972,594	1,381,107
Sequences identified by BLAST	790,424,528	574,477
Viral sequences identified by BLAST	91,863,018 (3.1%)	18,539 (1.3%)
Bleedover depleted viral sequences	91,654,946 (3.1%)	-
Bacterial sequences identified by BLAST	360,359,247 (12%)	411,889 (30%)
Other domain sequences identified by BLAST	338,202,263 (11%)	144,049 (10%)
Uncharacterized sequences	2,163,548,066 (73%)	806,628 (58%)

Viral sequences linked to laboratory components

From BLAST of reads and contigs 2994 viral clusters were identified (see Methods). Of these, significant associations were found between 493 viral clusters and laboratory components, hereafter referred to as laboratory component associated (LCA) viral sequences (see Fig. 1 and Supplementary material, Fig. S2A, Fig. S2B and Table S4). Remarkably, 68% (62,521,069) of all viral reads were included in viral clusters linked to laboratory components (see Supplementary material, Fig. S3A). For viral contigs this number was 74% (13,687 contigs). The majority of LCA viral sequences were non-human (see Supplementary material, Fig. S3B), with 60% (296/493) being bacteriophages.

Some of the laboratory components showed high correlation when investigating the extent of simultaneous use (see Supplementary material, Fig. S4), which can explain viral sequences showing significant association to multiple laboratory components. A particularly high proportion of viral sequences linked to laboratory components was seen for RNA-targeting overall methods (see Supplementary material, Fig. S3C). Components used as part of RNA methods (RNeasy MinElute, ScriptSeq v2, ScriptSeq Gold, RQ1 DNase and RQ1 Stop Solution) also showed the highest number of linked viral sequences (see Fig. 2).

In-silico verification of viral sequences linked to laboratory components

Mapping of reads to reference genomes was conducted to identify genome coverage and in-silico validate results from the BLAST analysis. Both global and local alignment was performed. The coverage of reference genomes was reported using the global mapping results, whereas local mapping was a complement used to confirm local BLAST hits. Cross-library genome coverage of reference genomes above 80% was seen for 13% (63/493) of LCA viral sequences (see Supplementary material, Table S5). Out of the 493 LCA viral sequences, 249 were linked to laboratory components based on global or local mapping of reads to reference genomes (see Supplementary material, Table S5). Detailed investigation of viral clusters showed viral sequences composed of sequences proposed to originate from the identified virus (referred to as true viral sequences) and/or viral sequences assumed to originate from an unknown or non-viral source (referred to as artefact viral sequences). The artefact viral sequences were short and regionally repeated nucleotide sequences, generally of low complexity or showing homology to cloning vectors or human sequences.

Human host viral sequences

In total, 24 LCA viral sequences from viruses having human as host were identified (see Table 3 and Supplementary material, Table S4 and Results S1). These viral sequences are particularly prone to erroneous conclusion when analysing human tissue samples. Low genome coverage (<25%) was identified in the majority of libraries (see Supplementary material, Fig. S5 and Table S6). A combination of sample-derived true viral sequences and laboratory component derived artefact sequences was identified for Human mast-adenovirus C, Human herpesvirus 1, Human herpesvirus 5, Human Immune-deficiency virus 1, Human parvovirus B19 and Torque teno virus. Among artefact sequences we identified homology to 1) various cloning or expression vectors (Human Immune-deficiency virus 1, Human parvovirus B19 and Semliki forest virus (see Supplementary material, Fig. S6 and Results S1)), 2) human sequences (Human papilloma-virus type 6), and 3) ribosomal RNA sequences (Simbu virus). Other artefact sequences did not show homology to specific types of sequences or were identified as short low complexity sequences.

193

194 **Table 3. Human viral sequences linked to laboratory components.** The table shows linked laboratory
195 components and *p*-value of association analysis based on BLAST identification, number of libraries in which
196 the viral sequence was identified by global mapping (N_{map}), distribution of alignments from the global
197 mapping, evaluation of origin (sample-derived and/or laboratory component derived (LCD)) and type of
198 sequence (true viral or artefact). For more detailed information see Supplementary table S4 and S5. *
199 (A/chicken/Karachi/NARC-100/2004(H7N3)). **(A/New York/55/2004(H3N2)). ***(B/Thailand/CU-
200 B2390/2010).

	Blast reads/contigs association analysis		**Mapping to reference genome**		
Viral sequence	**Linked laboratory component**	***P*-value**	**N_{map}**	**Alignments global mapping (% coverage)**	**Evaluation of sequence origin and identity**
Cyclovirus PK6197	RNeasy MinElute	2.5E-11	1	• 1 region 30bp (1.7%)	LCD unknown artefact
Coxsackievirus B1	RQ1 Stop Solution	8.8E-09	0	-	LCD unknown artefact
Human mast-adenovirus C	QIAamp DNA	2.7E-10	151	• Dispersed (0.14%-85%) • 1 region <40bp (<0.2%)	Sample-derived true viral sequences and LCD low complexity poly (A) artefact
Hepatitis E virus	ScriptSeq Gold	2.1E-12	0	-	LCD unknown artefact
Hepatitis C virus genotype 1	ScriptSeq v2	3.4E-19	537	• 4 regions <159bp (<2%)	LCD low complexity poly (T) artefact
Hepatitis C virus subtype 1b	ScriptSeq v2	9.5E-10	533	• 4 regions <277bp (<3%)	LCD low complexity poly (T) artefact
Human herpesvirus 1	PCR primers II	6.7E-11	81	• Dispersed (0.10%) • 3 regions <91bp (<0.06%)	Sample-derived true viral sequences and LCD unknown artefact
Human herpesvirus 4	ScriptSeq v2	2.5E-25	12	• 2 regions <42bp (0.03%)	LCD cloning vector artefact
Human herpesvirus 5	TURBO DNase	1.7E-21	43	• Dispersed (0.11%, 0.30%) • 1 region <208bp (<0.1%)	Sample derived true viral sequences and LCD cloning vector artefact
Human Immune-deficiency virus 1	ScriptSeq v2	2.8E-33	113	• Dispersed (11%-67%) • 3 regions <205bp (<3%)	Sample derived true viral sequences and LCD cloning vector artefact
Human papilloma-virus type 1a	Nextera XT	6.6E-10	16	• Dispersed (<4%)	True viral sequences of unknown origin
Human papilloma-virus type 6	PAXgene	2.0E-07	8	• Dispersed (0.39%, 1.4%) • 3 regions <32bp (<0.4%)	Human artefact
Human parvovirus B19	ScriptSeq v2	6.1E-12	41	• Dispersed (0.70%-100%) • 1 region <37bp (<0.7%)	Sample-derived true viral sequences and LCD expression vector artefact
Human T-lymphotropic virus 1	ScriptSeq v2	2.1E-31	81	• 1 region <151bp (<2%)	LCD unknown artefact
Influenza A virus*	Platinum Taq	2.8E-10	0	-	LCD unknown artefact
Influenza A virus**	ScriptSeq v2	1.2E-07	0	-	LCD unknown artefact
Influenza B virus***	Platinum Taq	1.2E-11	3	• Dispersed (1.5%) • 1 region <40bp (<2%)	LCD unknown artefact
Lassa virus	ScriptSeq v2	7.8E-46	0	-	LCD unknown artefact
Merkel cell polyomavirus	Nextera XT	6.9E-17	34	• Dispersed (0.76%-76%)	LCD true viral sequences

Macaca mulatta polyomavirus 1	ScriptSeq Gold	1.2E-13	23	• Dispersed (11%) • 2 regions <93bp (<2%)	Cross-mapping JC polyoma-virus and LCD expression and cloning vector artefact
Semliki Forest virus	ScriptSeq Gold	5.4E-10	0	-	LCD cloning vector artefact
Shamonda virus	RQ1 DNase	2.3E-11	0	-	LCD unknown artefact
Simbu virus	RQ1 Dnase	3.2E-09	167	• 1 region <49bp (0.40-0.70)	LCD ribosomal RNA artefact
Torque teno virus	Plasmid-Safe	1.9E-08	26	• Dispersed (3.2%-96%) • 1 region <351bp (<10%)	Sample derived true viral sequences and bleedover contamination

Non-human vertebrate host viral sequences

We identified 60 viral sequences from viruses with a non-human vertebrate host among LCA viral sequences (see Fig. 1 and Supplementary material, Fig. S2A and Table S4). Among these, 29 were avian retroviruses (predominantly from the *Alpharetrovirus* genus), also including Tasmanian devil retrovirus (see Supplementary material, Results S2). The avian retroviral sequences were linked to ScriptSeq v2 and/or ScriptSeq Gold and were identified in high proportions (median above 60%) in libraries prepared using these methods (see Fig. 3). The cross-library genome coverage ranged from 4-100% (see Supplementary material, Table S5 and Circos plots S1) and 17 viral sequences showed coverage above 60% with dispersed alignments, therefore proposed to be true viral sequences originating from laboratory components. Remaining avian retroviral sequences are considered artefact viral sequences and true viral sequences present at low quantities. From the *Parvovirinae* subfamily 13 viral sequences were identified. These viral sequences showed cross-library genome coverage of 95-100% and were all linked to the use of RNeasy MinElute, thus proposed to be laboratory component derived true viral sequences. Four viral sequences from the *Gammaretrovirus* genus linked to Nextera were identified. These showed regionally repeated alignments with relatively low cross-library genome coverage (7.8%-29%). The gammaretroviral sequences were, however, detected using several additional library preparation methods and we propose these sequences to be artefacts of unknown origin. In addition, 14 vertebrate viral sequences from eight different viral families were identified. Among these, Circovirus-like NI/2007-3 linked to RNeasy MinElute showed high cross-library genome coverage (96%), proposed to be a laboratory component derived true viral sequence. ASFV-like virus WU showed high cross-library genome coverage (80%) with dispersed alignments. It is considered a true viral sequence originating from laboratory components. The remaining vertebrate

LCA viral sequences showed low coverage (<3%) with no or regionally repeated alignments, indicating artefact viral sequences originating from laboratory components.

Furthermore, a high number of LCA viral sequences from viruses with non-vertebrate hosts were identified (see Fig. 1 and Supplementary material, Fig. S2A, Fig. S2B and Table S4). Among these were 25 algae host viral sequences, containing 14 chlorella viruses belonging to the *Chlorovirus* genus, including Acanthocystis Turfacea Chlorella virus (ATCV). All chlorella viral sequences were linked to RNeasy MinElute and showed no or low cross-library genome coverage (<5%) with dispersed and regionally repeated alignments (see Supplementary material, Table S5 and Circos plots S1). Chlorella viral sequences are proposed to be laboratory component derived artefact and true viral sequences present at low quantities. The remaining algae viral sequences showed no or low cross-library genome coverage (<2%) with dispersed as well as regionally repeated alignments, indicating both artefact and true viral sequences originating from laboratory components. In order to investigate if the presence of algae viral sequences could be explained by the presence of algae, reads were globally mapped to six algae chloroplast genomes. The observed cross-library genome coverage was 6.4-12% (see Supplementary material, Circos plots S2). BLASTn of the mapped reads against the complete NCBI nucleotide database identified the same chloroplast genomes, thereby supporting the presence of algae sequences in our libraries. Moreover, 18 invertebrate, 14 environmental, 3 fungal, 8 plant, 13 protozoan, 28 unknown and 296 bacterial host viral sequences were identified (see Supplementary material, Results S3).

Non-template controls

Eight of the LCA viral sequences were not detected in any of the NTCs (see Fig. 4A). Noteworthy, all eight were associated to the RNeasy MinElute kit. Among LCA viral sequences detected in the NTCs, the contaminating sequences were generally found in a higher proportion in NTC libraries than template containing libraries. We can estimate the power to successfully detect the virus from the frequency of each specific virus in the NTCs. Taking avian myeloblastosis virus and rodent stool-associated circular genome virus as examples; their respective detection frequencies in NTCs are 0.67 and 0.15. Assuming a binomial distribution, the probability of detecting these viral sequences if running three NTCs would be 0.96 and 0.39, respectively. In order to reach a

253 probability of detection higher than 0.95 for the rodent stool-associated circular genome virus, 19
254 NTCs would be necessary. Fig. 4B shows the number of NTCs needed for detection of a
255 contaminating viral sequence in one or more NTCs with a 95% probability, illustrating the
256 increasing number of NTCs necessary with decreasing detection rate.
257

258 **Discussion**

259 We have here provided a comprehensive list of 493 viral sequences, shown to be present in a
260 variety of sample types and NTCs, significantly associated to the use of one or more laboratory
261 components. Viral sequences showed stochastic appearance and were only detected in a subset of
262 the libraries treated with the linked laboratory component, not always appearing in the NTCs. The
263 host of linked viruses were taxonomically very diverse and included bacteria, protozoa, algae,
264 plants, fungi, invertebrates and vertebrates.

265 To our knowledge, this is the first study using a systematic approach to identify a
266 wide repertoire of contaminating viral sequences and their origin. Several laboratory protocols and
267 different laboratory components commonly used for sample preparation in virus discovery and
268 surveillance with HTS were applied to the same samples, which facilitated the identification of
269 laboratory component derived viral sequences. This is in contrast to other HTS studies where one
270 laboratory practice has been applied to samples (28).

271 Viral sequence contamination in clinical samples is a known occurrence. Several
272 viruses first linked to disease (19,21) have later been refuted as contamination (11,13,29,30). In
273 2014, Yolken et al. linked ATCV-1 to altered cognitive function after its detection in the throat of
274 healthy humans (20,31). Subsequently, we refuted these findings and suggested that ATCV-1
275 corresponded to contamination arising from one or two laboratory components used concurrently
276 during library preparation (17). In 2014, a pipeline for identifying pathogens in HTS data from
277 clinical samples was presented and applied to eight already published datasets originating from
278 samples of various disease origins (32). Noteworthy, 7.8% (76/974) of the viral findings in the
279 study from 2014 were in this study linked to laboratory components.

Among identified viral sequences many lacked significant association to laboratory components. Many of these viruses are non-human and identification of these would not be expected in human cancer tissue samples. Some of the viral sequences have previously been suggested to be contamination, such as Pepino mosaic virus (33), Gallid herpes virus, Pandoravirus, Citrobacter phage (34) and Rotavirus (32,35,36).

More viral sequences significantly associated to laboratory components were identified among reads than contigs, a probable explanation being the low depth of coverage of genomes, making assembly of viral reads into contigs infrequent. Viral contigs were therefore only identified in a subset of libraries and more frequently when applying enrichment methods. This is a limitation of our study. As enrichment methods are expected to be better for detection of particularly low amounts of viral sequences, certain contaminating viral sequences may be statistically linked to enrichment-specific laboratory components because of their ability to detect them, though potentially having a different source. A notorious problem in HTS are reads occurring as a result of mixed sequencing clusters, bleedover. This phenomenon could explain the presence of specific viral sequences in NTCs (at ratios higher than the applied bleedover threshold). Our results indicate higher bleedover ratios in NTCs than in template containing samples. Non-stringent e-values as low as 10^{-3} were applied in this study to reflect what is sometimes being applied in virus detection studies (37). However, more stringent e-values are necessary (and are being applied) when using HTS to diagnose viral infection (38).

In regard to detection of novel contaminating viral sequences, we rely on some degree of sequence similarity in the BLAST identification. This is a limitation of the analysis and further effort could be put into identifying contaminating viral sequences in the unidentified sequences. A sequence recurrence-based clustering method has recently been published (15). The strength of this approach is the independence of a sequence reference database for identification of correlation between nucleotide sequences and sample features.

Several viral sequences with significant association to several laboratory components were identified. Some of the laboratory components were used in parallel or almost in parallel, which resulted in identical or similar *p*-values, making it hard to be certain of the origin.

The factual origin(s) of viral sequences could be verified by setting up a designed experiment with different combinations of laboratory components and using virus specific PCR. This was however beyond the scope of this study.

Independent of the bioinformatic method used for the initial identification of viral sequences, we find it of outmost importance to evaluate genome coverage, read depth and the distribution of alignments across the identified viral genome. A high coverage and/or dispersed sequences across the reference genome indicate that the viral sequences are derived from the virus in question rather than representing an artefact. Short regionally repeated viral sequences in multiple samples indicate artefact viral sequences and should always raise suspicion. Re-blasting of regionally repeated LCA viral sequences showed additional best hits to cloning and expression vectors, for example, Semliki Forest virus, used as a vector for vaccine development, for gene therapy and for production of recombinant proteins (39-41). Others showed no additional best hit and the artefact sequence could therefore not be identified. The wide use of viral cloning and expression vectors could be an overlooked problem in virus discovery, leading to false positives.

Concerning the suitability of NTCs, our main conclusion is that several negative controls should be included in order to detect sporadic contaminants, despite the costs of sequencing (see Supplementary material, Discussion S1). Furthermore, we strongly recommend that viral sequences from viruses with non-human hosts should be handled with caution when identified in HTS data and that researchers carefully consider the possibility of contamination. It should be noted that viral sequences which we propose as contamination in human tissue samples, could have a natural origin in samples from another species or environmental location.

Supplementary material

Supplementary material is available on request from authors or can be downloaded from http://www.cbs.dtu.dk/public/contamination/.

Transparency declaration

The authors declare no conflicts of interest. This work was supported by Innovation Fund Denmark grant No 019-2011-2 (The Genome Denmark platform) and Danish National Research Foundation grant No DNRF94. The funders had no role in study design, data collection, analysis and interpretation, decision to submit the work for publication, or preparation of the manuscript.

Parts of the data have previously been presented in Geneva October 2018 at the International Conference on Clinical Metagenomics (ICCMg).

MA and KRK wrote the manuscript. SM and JMGI made major revisions of the manuscript. Laboratory experiments were designed by SM, HF, LV, KRK and RHJ and performed by SM, KRK, HF, LV, IBN, SRR, RHJ, ARI, DEAP and PVSO. Study design was done by AJH, JMGI, KRK, MA, TM, SM, LPN, OL, SB, TSP and EW. MA, JFN, JMGI and TAH designed bioinformatic pipeline. Initial bioinformatic analysis (pre-processing, assembly and BLAST) was conducted by JFN, JMGI and MLMH. MA and KRK performed the concluding analysis (clustering, data mining, statistical analysis, visualization). Mapping analysis and creation of Circos plots was done by JARH.

We thank BGI Europe for sequencing of the samples.

355 **References**

356 1. Reuter JA, Spacek DV, Snyder MP. High-throughput sequencing technologies. Mol Cell.
357 Elsevier; 2015 May 21;58(4):586–97.

358 2. Oulas A, Pavloudi C, Polymenakou P, Pavlopoulos GA, Papanikolaou N, Kotoulas G, et al.
359 Metagenomics: tools and insights for analyzing next-generation sequencing data derived from
360 biodiversity studies. Bioinformatics and Biology Insights. 2015;9:75–88.

361 3. Briese T, Paweska JT, McMullan LK, Hutchison SK, Street C, Palacios G, et al. Genetic
362 detection and characterization of Lujo virus, a new hemorrhagic fever-associated arenavirus
363 from southern Africa. PLoS pathogens. 2009 May;5(5):e1000455.

364 4. Yozwiak NL, Skewes-Cox P, Gordon A, Saborio S, Kuan G, Balmaseda A, et al. Human
365 enterovirus 109: a novel interspecies recombinant enterovirus isolated from a case of acute
366 pediatric respiratory illness in Nicaragua. Journal of virology. 2010 Sep;84(18):9047–58.

367 5. Feng H, Shuda M, Chang Y, Moore PS. Clonal integration of a polyomavirus in human
368 Merkel cell carcinoma. Science (New York, NY). 2008 Feb;319(5866):1096–100.

369 6. Cholleti H, Hayer J, Abilio AP, Mulandane FC, Verner-Carlsson J, Falk KI, et al. Discovery of
370 Novel Viruses in Mosquitoes from the Zambezi Valley of Mozambique. Datta S, editor. PloS
371 one. 2016;11(9):e0162751.

372 7. Hansen TA, Fridholm H, Frøslev TG, Kjartansdóttir KR, Willerslev E, Nielsen LP, et al. New
373 Type of Papillomavirus and Novel Circular Single Stranded DNA Virus Discovered in Urban
374 Rattus norvegicus Using Circular DNA Enrichment and Metagenomics. Angeletti PC, editor.
375 PloS one. Public Library of Science; 2015;10(11):e0141952.

376 8. Ng TFF, Manire C, Borrowman K, Langer T, Ehrhart L, Breitbart M. Discovery of a novel
377 single-stranded DNA virus from a sea turtle fibropapilloma by using viral metagenomics.
378 Journal of virology. American Society for Microbiology; 2009 Mar;83(6):2500–9.

9. Munang'andu HM, Mugimba KK, Byarugaba DK, Mutoloki S, Evensen Ø. Current Advances on Virus Discovery and Diagnostic Role of Viral Metagenomics in Aquatic Organisms. Front Microbiol. Frontiers; 2017;8(Pt 6):406.

10. Chiu CY. Viral pathogen discovery. Current opinion in microbiology. 2013 Aug;16(4):468–78.

11. Smuts H, Kew M, Khan A, Korsman S. Novel hybrid parvovirus-like virus, NIH-CQV/PHV, contaminants in silica column-based nucleic acid extraction kits. Journal of virology. 2014 Jan;88(2):1398–8.

12. Lusk RW. Diverse and widespread contamination evident in the unmapped depths of high throughput sequencing data. PloS one. 2014;9(10):e110808.

13. Naccache SN, Greninger AL, Lee D, Coffey LL, Phan T, Rein-Weston A, et al. The Perils of Pathogen Discovery: Origin of a Novel Parvovirus-Like Hybrid Genome Traced to Nucleic Acid Extraction Spin Columns. Journal of virology. 2013 Oct;87(22):11966–77.

14. Lysholm F, Wetterbom A, Lindau C, Darban H, Bjerkner A, Fahlander K, et al. Characterization of the viral microbiome in patients with severe lower respiratory tract infections, using metagenomic sequencing. Highlander SK, editor. PloS one. Public Library of Science; 2012;7(2):e30875.

15. Friis-Nielsen J, Kjartansdóttir KR, Mollerup S, Asplund M, Mourier T, Jensen RH, et al. Identification of Known and Novel Recurrent Viral Sequences in Data from Multiple Patients and Multiple Cancers. Viruses. Multidisciplinary Digital Publishing Institute; 2016 Feb 19;8(2):53.

16. Laurence M, Hatzis C, Brash DE. Common Contaminants in Next-Generation Sequencing That Hinder Discovery of Low-Abundance Microbes. PloS one. 2014 May;9(5):e97876–8.

17. Kjartansdóttir KR, Friis-Nielsen J, Asplund M, Mollerup S, Mourier T, Jensen RH, et al.

Traces of ATCV-1 associated with laboratory component contamination. Proceedings of the National Academy of Sciences of the United States of America. National Acad Sciences; 2015 Mar 3;112(9):E925–6.

18. Lo S-C, Pripuzova N, Li B, Komaroff AL, Hung G-C, Wang R, et al. Detection of MLV-related virus gene sequences in blood of patients with chronic fatigue syndrome and healthy blood donors. Proceedings of the National Academy of Sciences of the United States of America. National Acad Sciences; 2010 Sep 7;107(36):15874–9.

19. Lombardi VC, Ruscetti FW, Gupta Das J, Pfost MA, Hagen KS, Peterson DL, et al. Detection of an infectious retrovirus, XMRV, in blood cells of patients with chronic fatigue syndrome. Science (New York, NY). 2009 Oct;326(5952):585–9.

20. Yolken RH, Jones-Brando L, Dunigan DD, Kannan G, Dickerson F, Severance E, et al. Chlorovirus ATCV-1 is part of the human oropharyngeal virome and is associated with changes in cognitive functions in humans and mice. Proceedings of the National Academy of Sciences of the United States of America. National Acad Sciences; 2014 Nov 11;111(45):16106–11.

21. Xu B, Zhi N, Hu G, Wan Z, Zheng X, Liu X, et al. Hybrid DNA virus in Chinese patients with seronegative hepatitis discovered by deep sequencing. Proceedings of the National Academy of Sciences of the United States of America. 2013 Jun;110(25):10264–9.

22. Schlaberg R, Choe DJ, Brown KR, Thaker HM, Singh IR. XMRV is present in malignant prostatic epithelium and is associated with prostate cancer, especially high-grade tumors. Proceedings of the National Academy of Sciences of the United States of America. 2009 Sep 22;106(38):16351–6.

23. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research. 1997 Sep;25(17):3389–402.

24. Kircher M, Sawyer S, Meyer M. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic Acids Research. Oxford University Press; 2012 Jan;40(1):e3–e3.

25. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria. Available from: https://www.R-project.org/

26. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012 Mar 4;9(4):357–9.

27. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an information aesthetic for comparative genomics. Genome research. 2009 Sep;19(9):1639–45.

28. Moustafa A, Xie C, Kirkness E, Biggs W, Wong E, Turpaz Y, et al. The blood DNA virome in 8,000 humans. Belshaw R, editor. PLoS pathogens. 2017 Mar;13(3):e1006292.

29. Erlwein O, Robinson MJ, Dustan S, Weber J, Kaye S, McClure MO. DNA extraction columns contaminated with murine sequences. Jeang KT, editor. PloS one. Public Library of Science; 2011;6(8):e23484.

30. Paprotka T, Delviks-Frankenberry KA, Cingöz O, Martinez A, Kung H-J, Tepper CG, et al. Recombinant origin of the retrovirus XMRV. Science (New York, NY). 2011 Jul;333(6038):97–101.

31. Yolken RH, Jones-Brando L, Dunigan DD, Kannan G, Dickerson F, Severance E, et al. Reply to Kjartansdóttir et al.: Chlorovirus ATCV-1 findings not explained by contamination. Proceedings of the National Academy of Sciences of the United States of America. National Acad Sciences; 2015 Mar 3;112(9):E927–7.

32. Naccache SN, Federman S, Veeraraghavan N, Zaharia M, Lee D, Samayoa E, et al. A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. Genome research. 2014 Jul;24(7):1180–92.

33. Bukowska-Ośko I, Perlejewski K, Nakamura S, Motooka D, Stokowy T, Kosińska J, et al. Sensitivity of Next-Generation Sequencing Metagenomic Analysis for Detection of RNA and DNA Viruses in Cerebrospinal Fluid: The Confounding Effect of Background Contamination. Advances in experimental medicine and biology. 2016 Jul;(Chapter 42):1–10.

34. Hjelmsø MH, Hellmér M, Fernandez-Cassi X, Timoneda N, Lukjancenko O, Seidel M, et al. Evaluation of Methods for the Concentration and Extraction of Viruses from Sewage in the Context of Metagenomic Sequencing. Tang P, editor. PloS one. Public Library of Science; 2017;12(1):e0170199.

35. Grard G, Fair JN, Lee D, Slikas E, Steffen I, Muyembe J-J, et al. A novel rhabdovirus associated with acute hemorrhagic fever in central Africa. PLoS pathogens. 2012 Sep;8(9):e1002924.

36. Grard G, Fair JN, Lee D, Slikas E, Steffen I, Muyembe J-J, et al. Correction: A Novel Rhabdovirus Associated with Acute Hemorrhagic Fever in Central Africa. PLoS pathogens. Public Library of Science; 2016 Mar;12(3):e1005503.

37. Prachayangprecha S, Schapendonk CME, Koopmans MP, Osterhaus ADME, Schürch AC, Pas SD, et al. Exploring the potential of next-generation sequencing in detection of respiratory viruses. Caliendo AM, editor. J Clin Microbiol. American Society for Microbiology Journals; 2014 Oct;52(10):3722–30.

38. Thézé J, Li T, Plessis du L, Bouquet J, Kraemer MUG, Somasekar S, et al. Genomic Epidemiology Reconstructs the Introduction and Spread of Zika Virus in Central America and Mexico. Cell Host Microbe. 2018 Jun 13;23(6):855–7.

39. Atkins GJ, Fleeton MN, Sheahan BJ. Therapeutic and prophylactic applications of alphavirus vectors. Expert reviews in molecular medicine. 2008;10:e33.

40. Lundstrom K. Alphavirus vectors as tools in neuroscience and gene therapy. Virus research.

2016 May;216:16–25.

41. DiCiommo DP, Bremner R. Rapid, high level protein production using DNA-based Semliki Forest virus vectors. The Journal of biological chemistry. 1998 Jul;273(29):18060–6.

Figure legends

Fig. 1. *P*-values of association analysis between eukaryotic viral sequences and laboratory components at contig level.

Including viral sequences linked to one or more laboratory components (e-value<10^{-3}). Significant associations illustrated in red and non-significant associations illustrated in blue. The strongest association(s) for each viral sequence is marked with a black star and white stars shows multiple-source associations. Laboratory components with minimum one linked viral sequence are marked in bold font.

Fig. 2. Number of viral sequences linked to the various laboratory components.

Counts comprise the number of viral sequences linked (showing the strongest association) to each laboratory component, including viral sequences linked to more than one laboratory component because of identical *p*-values, and multiple source viral sequences. Bars to the left and right show results from BLAST of reads and contigs, respectively.

Fig. 3. **Avian retroviral read ratios for the different library preparation methods**.

The black lines illustrate the median and the red dots illustrate the average ratio of avian retroviral reads for the different library preparation methods.

Fig. 4. Non-template controls.

(A) Detection rates of specific LCA viral sequences in NTC libraries and template containing libraries. The number of NTC libraries and template containing libraries is shown above the bars in the figure. (B) Number of NTCs necessary to reach a detection probability of minimum 0.95 for different viral detection rates. Legend: detection rate : number of NTCs.

Supplementary material legends

Fig. S1. Laboratory components applied to samples.

Laboratory components applied to samples before HTS. Laboratory components ordered according to category and libraries ordered according to overall method. A) DNA methods. B) RNA methods.

Fig. S2. P-values of association analysis between viral sequences and laboratory components.

Including viral sequences linked to one or more laboratory components (e-value<10^{-3}). Significant associations illustrated in red and non-significant associations illustrated in blue. The strongest association(s)

for each viral sequence is marked with a black star and white stars shows multiple-source associations. Laboratory components with significant association to minimum one viral sequence are marked in bold font. (A) Association analysis of viral reads. (B) Association analysis of viral contigs.

Fig. S3. Allocation of sequences.

(A) BLAST characterization of reads (human and low complexity depleted) and contigs into different groups and association analysis of reads and contigs characterized as viral sequences. (B) Number of LCA viruses with different hosts; reads and contigs. (C) Ratio of reads associated to laboratory components among all characterized reads (red boxes) and viral reads (blue boxes) for different overall methods (grey boxes above boxplot) and different library preparation methods (y-axes); black line illustrating the median value and red dot the average value among libraries.

Fig. S4. Correlation between laboratory components.

Each square showing the Pearson correlation (r) between laboratory components (ordered according to laboratory component category) with colour indicating the strength of the correlation; scaling to the right. Names of laboratory components with minimum one correlation above 0.9 are marked in red and those with minimum one complete correlation (r=1) in dark red.

Fig. S5. Coverage of human LCA viruses.

Number of libraries with specific coverage from the global mapping of reads to the selected reference genomes.

Fig. S6. Semliki Forest virus.

Mapping of contigs (red) to the Semliki Forest virus genome (Z48163.2) including the cloning vector (blue).

Table S1. Sequencing libraries.

Table includes library sequenced, sample processed, sample type, overall method for library preparation, sequencing lane identification number, number of reads before (total read count) and after depletion (human and low complexity depletion) (read count depleted), number of assembled contigs, number of characterized sequences (hit count) for reads and contigs and number of viral sequences identified (viral hit count).

Table S2. Laboratory components.

Laboratory components investigated for correlation to viral sequences.

Table S3. Selected reference genomes for LCA viruses.

Table includes taxID description and taxID identification number of viral sequences. For each selected reference genome; accession and GI number, GI definition, database from which sequence was recovered and length in nucleotides (nt).

Table S4. Viral sequences linked to laboratory components.

Viral sequences with a significant association to a laboratory component. Table includes taxID description and taxID identification number of viral sequences, host, separation of viral sequences into categories, parental taxa, I a) BLASTn of reads; total number of reads within viral cluster (across all libraries), mean and median percent identity, mean and median alignment length, mean and median e-value, number of reads within viral cluster (across all libraries) after bleedover removal, number of libraries within cluster, mean and median read count and mean and median read count proportion (read count divided with the number of viral sequences identified in each library) of libraries within cluster, II a) BLASTnx of contigs; number of assembled contigs within viral cluster, number of libraries within cluster, mean and median percent identity, mean and median alignment length and mean and median e-value, I b) Association analysis based on BLAST characterized reads and II b) Association analysis based on BLAST characterized contigs; strongest associated feature (F), percentage of libraries having applied F where viral sequence was detected (LC+ det. %), number of libraries having applied F where viral sequence was detected (LC+ det. count), percentage of libraries not having applied F where viral sequence was detected (LC- det. %), number of libraries not having applied F where viral sequence was detected (LC- det. count), p-value of association to F.

Table S5. Cross-library mapping results for LCA viruses.

Table includes taxID description and taxID identification number of viral sequences, host, category and accession number of selected reference genome. I) Global mapping of reads to the selected reference genome and II) Local mapping of reads to the selected reference genome; cross-library genome coverage (%) for all libraries, average coverage percent (%) per library, average read depth per library, average read count of reads mapping per library, numbers of libraries where viral sequences mapped (excluding bleedover), number of libraries where viral sequences mapped determined as bleedover. III) Association analysis based on global mapping of reads to the selected reference genome, IV) Association analysis based on local mapping of reads to the selected reference genome; strongest associated feature (F), percentage of libraries having applied F where reads mapped (LC+ det. %), number of libraries having applied F where reads mapped (LC+ det. count), percentage of libraries not having applied F where reads mapped (LC- det.

%), number of libraries not having applied F where reads mapped (LC- det. count), p-value of association to F.

Table S6. Library specific mapping results for LCA viruses.

Table includes taxID description and taxID identification number of viral sequences and accession number of selected reference genome. Library information; library number, sequencing lane number, sample processed, sample type, overall method used to prepare library and library preparation method. I) Global mapping of reads to the selected reference genome and II) Local mapping of reads to the selected reference genome; coverage percent (%), coverage in nucleotides (nt), read depth, read count of reads mapping, the highest number of reads mapping to the same reference genome from a library sequenced on the same lane (read count max) and calculated bleedover ratio.

Circos plots S1. Coverage of LCA viruses.

Visualization of global mapping of reads to selected reference genomes of LCA viruses.

Circos plots S2. Coverage of algae chloroplast genomes.

Visualization of global mapping of reads to six selected algae chloroplast genomes.

Laboratory methods S1

Results S1. Human viral sequences identified as contaminants.

Results S2. Tasmanian devil retrovirus included among avian retroviruses.

Results S3. Non-vertebrate viral sequences linked to laboratory components

Discussion S1.