High-throughput sequencing-based investigation of viruses in human cancers by multi-enrichment approach

Mollerup, Sarah; Asplund, Maria; Friis-Nielsen, Jens; Kjartansdóttir, Kristín Rós; Fridholm, Helena; Hansen, Thomas Arn; Herrera, José Alejandro Romero; Barnes, Christopher James; Jensen, Randi Holm; Richter, Stine Raith

Published in:
Journal of Infectious Diseases

Link to article, DOI:
10.1093/infdis/jiz318

Publication date:
2019

Document Version
Peer reviewed version

Link back to DTU Orbit

Citation (APA):
Title
High-throughput sequencing-based investigation of viruses in human cancers by multi-enrichment approach

Authors

*Corresponding authors, email: sarahmollerup@gmail.com, ajhansen@snm.ku.dk

Correspondence during submission/review process to: Sarah Mollerup, Oester voldgade 5-7, Dk-1550 Copenhagen K, Denmark. Phone: +45 40516216

# These authors contributed equally to this article.

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

© The Author(s) 2019. Published by Oxford University Press for the Infectious Diseases Society of America.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
2 Department of Bio and Health Informatics, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark.

3 Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen DK-2200, Denmark.

4 Department of Growth and Reproduction, Copenhagen University Hospital (Rigshospitalet), 9 Blegdamsvej, DK-2100 Copenhagen, Denmark.

5 Department of Pediatrics and Adolescent Medicine, University Hospital Rigshospitalet, Copenhagen, Denmark.

6 Department of Otorhinolaryngology, Head and Neck Surgery and Audiology, Rigshospitalet, Copenhagen University Hospital.

7 Department of Dermato-Venerology, Faculty of Health Sciences, Copenhagen University Hospital, Bispebjerg Hospital, Denmark.

8 Department of Pathology, Herlev and Gentofte Hospital, University of Copenhagen, Denmark.

9 National Institute of Oncology, Department of Gynecology, Budapest, Hungary.

10 1st Department of Surgery, Semmelweis University, Budapest, Hungary.

11 Department of Clinical Medicine, Aalborg University, Aalborg, Denmark.

12 Department of Haematology, Aalborg University Hospital, Aalborg, Denmark.

13 Department of Pathology, Aarhus University Hospital, Aarhus, Denmark.

14 Department of Clinical Medicine, Department of Haematology, Aarhus University Hospital, Aarhus, Denmark.

15 Department of Surgery, Herlev and Gentofte Hospital, University of Copenhagen, Denmark.

16 Center for Clinical Research, North Denmark Regional Hospital and Department of Clinical Medicine, Aalborg University, DK-9800 Hjørring, Denmark.

17 Centre of Excellence for Omics-Driven Computational Biodiscovery, AIMST University, Kedah, Malaysia.
Summary of main points

High-throughput sequencing of nearly 200 cancer samples detected viruses from seven viral families. More than half of the investigated samples contained one or more viruses, however, no associations linking specific viruses with specific cancer types were found.
Abstract
Viruses and other infectious agents cause more than 15% of human cancer cases. High-throughput sequencing-based studies of virus-cancer associations have mainly focused on cancer transcriptome data. Here we applied a diverse selection of pre-sequencing enrichment methods targeting all major viral groups, to characterise the viruses present in 197 samples from 18 sample types of cancerous origin. Using high-throughput sequencing we generated 710 datasets constituting 57 billion sequencing reads. Detailed in silico investigation of the viral content, including exclusion of viral artefacts, from de novo assembled contigs and individual sequencing reads yielded a map of the viruses detected. Our data reveals a virome dominated by papillomaviruses, anelloviruses, herpesviruses, and parvoviruses. More than half of the included samples contained one or more viruses, however, no link between specific viruses and cancer types were found. Our study sheds light on viral presence in cancers and provides highly relevant virome data for future reference.

Keywords
Cancer, virus, virome, high-throughput sequencing, sequencing, next generation sequencing, human, enrichment, in-depth analysis, artefacts
Background

Globally, more than 15% of human cancer cases occurring in 2008 could be ascribed to infectious agents classified as carcinogenic according to the International Agency for Research on Cancer (IARC) [1]. This excludes viruses and cancer sites for which evidence of carcinogenicity is weaker. The IARC-classified carcinogenic agents include six types of viruses: hepatitis B and C virus, high-risk human papillomaviruses (HPVs), human herpesvirus (HHV) 4 (Epstein-Barr virus), human T-cell lymphotropic virus type, and HHV 8 (Kaposi’s sarcoma-associated herpesvirus). Hepatitis B virus-associated hepatocellular carcinoma and human papillomavirus-associated cervical and anal cancer can be prevented through vaccination [2],[3]. Apart from both firmly and less firmly established associations, additional cancers might be caused by either known or unknown viruses and could therefore be preventable.

With the introduction of high-throughput sequencing, description of the virome of various tissues of both healthy and diseased individuals has accelerated [4],[5],[6],[7],[8],[9],[10],[11],[12],[13], generating important knowledge about the viral species hosted by humans. Application of high-throughput sequencing led to the discovery of Merkel cell polyomavirus suspected of causing Merkel cell carcinomas [14] and in later years, large scale investigations of viral expression in high-throughput RNA-sequencing data, and of viral sequences in whole genomes or exomes based on data from The Cancer Genome Atlas have been conducted [15],[16],[17]. These studies have confirmed established virus-cancer associations and raised questions about hypothesised associations, but have so far not revealed novel associations.

Infection with carcinogenic viruses is common, but only rarely leads to cancer. Upon transformation, the virus persists intracellularly as an episome, or integrated in the host cell genome [18]. To target the multiple possible types and stages of viral genomes we applied sensitive pre-sequencing methods for enrichment of virions [19], circular DNA genomes [20], and for capturing retroviral [21] or vertebrate viral sequences [22]. The methods were applied, along with high-
throughput sequencing of total DNA and RNA, to 197 samples from 18 cancer types (including biopsies, bone marrow, and urine samples) as well as samples of ascites, blood from colon cancer patients, and a few healthy control samples. Targeting a breadth of viruses, our study presents a comprehensive characterisation of the virome of the included cancer samples, thus expanding the reference catalogue of the viruses found in these cancers.

Methods

Samples and datasets

The present study includes 760 datasets generated from 197 patient samples and 50 non-template controls. Some of the datasets were included in previous studies, see Supplementary Methods. Viral sequence contamination in the included samples is explored in detail in a separate study [23] (in press). The description of all samples, laboratory, and bioinformatic methods applied are provided here for the sake of completeness.

Ethics statement

Human sample collection, handling, and analysis were performed under the ethical protocols H-2-2012-FSP2 (Regional Committee on Health Research Ethics) and case no. 1304226 (National Committee on Health Research Ethics). In accordance with National legislation (Sundhedsloven), all human samples were processed anonymously.

Patient samples

All samples are listed in Table 1. Detailed information regarding samples and datasets can be found in Supplementary Methods and Table S1.
Total DNA analysis

Total DNA was extracted using the QIAamp DNA Mini kit (Qiagen). DNA libraries were prepared from 1 μg of DNA using either the TruSeq DNA protocol (PE-940-2001) (Illumina) or an in-house protocol [24] using NEBNext reagents (E6070) (New England BioLabs).

Total RNA and mRNA analysis

Total RNA was extracted using the High Pure Viral RNA kit (Roche), RNeasy Mini Kit (Qiagen), or QIAamp DNA Mini Kit (Qiagen). mRNA was extracted using Dynabeads mRNA Direct Purification Kit (Invitrogen). RNA libraries were prepared using ScriptSeq v2 RNA-Seq or ScriptSeq Complete Gold Library Preparation Kit (Epicentre). See Supplementary Methods for details regarding extraction kits, rRNA depletion, and library preparation kits used.

Circular DNA enrichment

Enrichment of small circular DNA molecules was performed on total DNA extracts based on phi29 DNA polymerase-mediated amplification of exonuclease treated extracts as previously described with minor modifications [20]. 2 μg of DNA was fragmented using the Bioruptor NGS (Diagenode) to an average length of 300 bp. Libraries were prepared as described for total DNA analysis.

Retrovirus capture

Two versions of retrovirus capture were applied. Retrovirus capture v1 includes 118 retroviral reference sequences [21] (Table S2). Capture was performed with 1 μg of single indexed libraries prepared from total DNA or mRNA (see above) according to the SeqCap EZ library SR protocol (Roche, NimbleGen) (capture dataset numbers between s0001 and s1112, Table S1). Retrovirus capture v2 includes 98 retroviral reference sequences (Table S2). Capture was performed with 500 μg of double indexed libraries prepared from total DNA according to the MYcroarray MYbaits
protocol version 2.3.1 with some modifications according to protocol version 1.3.8 (separating the beads from the eluted captured library and addition of Neutralization buffer to the supernatant) (capture dataset numbers s1431-s1440, Table S1).

Vertebrate virus capture DNA

The vertebrate virus capture probe design includes 2339 sequences representing viral species found in vertebrates, excluding fish [22] (Table S2). Sequences representing Merkel cell polyomavirus, KI polyomavirus, and HHV5 were not included in genomes used for probe design. SeqCap EZ hybridization probes were designed and synthesized by Roche NimbleGen (Madison, USA). Capture was performed on double indexed libraries prepared from total DNA extracted using DNeasy Blood and Tissue (Qiagen) or QIAamp DNA Mini kit (Qiagen). Libraries were prepared as described for total DNA analysis. Viral sequences were captured from 1 μg of pooled libraries as described in [22] with the following modifications: hybridization buffer without 10% formamide was used, and the amplified captured libraries were purified using QIAquick PCR Purification Kit (Qiagen).

Enrichment of virion-associated DNA and RNA

Samples used for enrichment were fresh frozen after collection with no addition of nucleic acid preservers. Enrichment was performed as previously described [25]. DNA libraries were prepared using the Nextera or Nextera XT DNA Sample Preparation Kit (Illumina) and RNA libraries using ScriptSeq v2 RNA-Seq Library Preparation Kit (Epicentre) and purified using the Agencourt AMPure XP PCR purification system (Beckman Coulter). In cases of insufficient amplification, libraries were re-amplified using AccuPrime Pfx DNA polymerase (Life Technologies) and P5 and P7 primers.
Sequencing and data analysis

Paired-end sequencing (2×100 bp) was performed on the Illumina HiSeq 2000 platform. The sequence analysis is detailed in the Supplementary Methods. Briefly, reads were trimmed of adapter sequences and overlapping read pairs merged. Human sequences were depleted by mapping to the human genome, and low complexity regions filtered out. De novo assembly was achieved using IDBA [26]. The reads and contigs were aligned to the NCBI nucleotide database (nt) using BLASTn (megablast) [27] with a cut-off e-value of 10^{-3}. The best hit was defined based on highest bit-score. Regions in the contigs having no BLASTn hits were aligned against the NCBI non-redundant protein database (nr) using BLASTx or DIAMOND [28] with a cut-off e-value of 10^{-3}. Individual reads for s1431-s1523 were not blasted.

Investigation of human viral hits

To exclude false positives all BLAST/DIAMOND hits to human viruses were evaluated in silico and categorised as confirmed viral hits or artefacts (see Supplementary Methods). For the contigs, hits were evaluated manually by alignment using Geneious software or web-based re-blast. For the reads, hits were evaluated by mapping to a database of 343 selected viral reference genomes. The alignments were visualised using Circos [29]. All plots were visually inspected. Hits arising from bleedover [30],[31] were removed from both mapping results and contigs. For the read mapping, a lower cut-off of 180 (205 for human immunodeficiency virus) bases covered was applied.

Co-occurrence network

Co-occurrence patterns between species occurring in four or more samples were investigated by performing Spearman’s rank correlations and network inference on the read mapping data. Human papillomaviruses and anelloviruses unclassified at species level were evaluated at strain level. Such strains, occurring in fewer than four samples were disregarded as well, leaving only two anellovirus strains unclassified at species level (here termed Unclassified Anellovirus 1 and 2). Non-template
controls were also excluded. Correlations were performed in vegan [32], while the network was constructed using igraph [33]. Networks were visualized using Cytoscape (v.3.6.0) [34].

Statistics

Comparison of the proportions of virus-positive samples was performed using Fisher’s exact test, with a significance level of 0.05. For the co-occurrence network, co-occurrences were considered significant when Spearman’s correlation coefficient was > 0.20 (P-value < 0.05) [35].

Data Availability

Sequencing data depleted of human sequences is deposited at the NCBI sequence read archive, Bioproject accession PRJNA416252. According to Danish law, publication of human sequences is not permitted without consent, which cannot be obtained, as all samples were anonymized. The complete coding sequences of HPV strain CGG5-287s1382c000001 and HPV strain CGG5-301s0532c000007 and six contigs representing shorter genome fragments of novel HPV types are uploaded to GenBank, accession numbers MG869604-MG869611.

Results

Investigation of human viral hits

We applied multiple viral enrichment methods (Fig 1) to 197 samples of diverse cancer types (Table S1), resulting in 710 datasets (Table 1) and 50 non-template (negative) controls constituting >57 billion Illumina HiSeq read pairs, with the median number of reads per dataset ranging from 30.5 to 169 million, depending on the method applied (Table S3). De novo assembly of the non-human fraction of the reads yielded ~1.5 million contigs. The taxonomy of contigs and reads was assigned using a BLAST-based pipeline (Fig 1 and Supplementary Material). These analyses are hereafter referred to as BLASTnx (for contigs) or BLASTn (for reads).
Investigation of the viral BLAST hits (Table S4) revealed artefacts arising mainly due to short local-only sequence similarity to viral genomes. All hits to human viruses were therefore evaluated in silico (see Supplementary Methods and Results). For the contigs, confirmed hits to 61 viruses from 6 viral families were found, while 14 human viruses were disregarded as false positives (Table S4). For the reads, mapping to 343 manually selected viral genomes, hereafter referred to as read mapping, confirmed viral hits to 146 reference genomes (Table S5 and for mapping results see Supplementary Data 1 and coverage plots in Fig S1). The artefactual viral sequences identified in our data are explored further in a separate study [23] (in press). Confirmed viral hits (Table S6 and Supplementary Data 2) were furthermore depleted of bleedover of viral reads occurring during sequencing.

The virome of the cancerous samples
Of the 197 samples included, 54 (27%) were virus-positive at contig level, while 106 (54%) were virus-positive from read mapping. For several skin-associated and mucosal cancer types, all samples were found virus-positive (Table S7), while certain sample types revealed no confirmed viral sequences. The detected viruses mainly belong to the families Papillomaviridae, Polyomaviridae, Herpesviridae, Parvoviridae, and Anelloviridae. Throughout the results, the identified viruses are grouped at species or genus level for both contig BLASTnx and read mapping (Fig 2), while the individual viral strains identified are presented fully in the Supplementary Material (Fig S2-S3). Between 2 and 7 different viral genera were represented in the virus-positive samples (median of 2) (4), with the highest diversity of viral genera generally occurring in skin-associated and mucosal cancers (Table S9).

Papillomaviruses
HPVs were detected mainly in skin and mucosa-associated cancers (64-73% of samples) (Table 2, Fig 2-3). De novo assembly recovered the full genome of a novel type of Gammapapillomavirus in
a single contig in an oral cavity cancer sample (HPV strain CGG5-301s0532c000007, Table S10), being most similar to HPV146 (Fig S5 and Supplementary Methods). Contigs representing shorter genome fragments of novel HPV types and full genomes of known types were also detected (Table S10). High-risk alphapapillomaviruses were found in a few samples; HPV16 and HPV18 in contigs (full genomes), and HPV18 and HPV42 from read mapping (at low coverage). HPV-positive skin-associated and mucosal samples contained sequences mapping to up to 17 different HPV types (median 2 types), with oral cavity cancers showing the highest numbers (median 5 types) (see Discussion). In skin-associated cancers, Betapapillomavirus was the most represented genus (Fig 2-3, Table S8), differing from previous studies of healthy skin [5],[9], while oral cavity cancers showed high both Beta- and Gammapapillomavirus positivity, also contrasting previous findings [9],[36].

Polyomaviruses

Polyomaviruses were detected mainly in skin-associated and certain mucosal cancers (Table 2, Table S8, Fig 2, Fig S2-S3). In bladder cancer urine, BK polyomavirus (BKV) (33-98% coverage, Table S6) and JC polyomavirus (JCV) (99% coverage) were detected, while most of the remaining polyomavirus-positive samples contained Merkel cell polyomavirus (MCPyV). One bladder cancer urine sample was found positive for both JCV (>10 million reads, 99% coverage) and BKV (59 reads, 8.4% coverage), the latter finding possibly arising due to sequence homology between these two viruses (see Supplementary Discussion). MCPyV was only detected when applying virion enrichment DNA, and SNPs were found to recur between positive datasets, suggesting a possible contamination (see Supplementary Discussion).

Herpesviruses

HHV1, HHV2, and HHV5 were detected in a few samples each, all at low coverage (up to 0.34%), while HHV4, 6, and 7 were more widespread (Fig 2, Table S8). HHV4 was found mainly in certain
skin and mucosa-associated cancers, while HHV6 was found mainly in malignant melanoma, and HHV7 mainly in bladder cancer, oral cavity cancer, and mycosis fungoides. HHV6B and HHV7 were of low coverage, except a sample of mycosis fungoides and testicular cancer showing higher HHV6A coverage (99% and 53%). HHV4 also showed higher genome coverage in certain samples (up to 69%). In all samples showing presence of HHV6A (Fig S3), reads mapping to both HHV6A and HHV6B were detected, likely arising due to sequence homology between these two species (see Supplementary Discussion).

Paroviruses

Human parovirus B19 was detected mainly in skin-associated cancers (80-91% of samples by read mapping, 32-100% coverage, Fig 2, Table S6 and S8). The recently described cutavirus of the genus Protoparvovirus [37] was detected from contigs and read mapping in one sample of malignant melanoma as presented earlier [38]. In addition, adeno-associated virus-2 was detected in a few samples.

Anelloviruses

Anelloviruses were detected in the contigs at highest prevalence in certain mucosal cancers and leukaemias (Table 2, Fig 2). Full or near full genomes were detected among the contigs (Table S11), some of these possibly representing novel anellovirus species. Contigs and reads mapping to different anelloviruses were often seen (Fig S2-S3), however species/strain-level identification of these might be less certain (see Discussion and Supplementary Material).

Rare occurrences

A few viruses occurred only sporadically. The flavivirus human pegivirus (formerly GB virus C) was detected in two samples of BCP-ALL, and one sample each of T-ALL, AML, and CML (2.1-
17% coverage Table S6), while HIV-1 was detected in ascites from a colon cancer patient (11% coverage) (Fig 2, Fig S3).

Co-occurrence of viruses

The non-random patterns of viruses detected in the different sample types were explored by investigation of co-occurrence of viruses. For this analysis viruses were grouped at species level, and only species identified in at least four samples by read mapping were included (Fig 4). Viral species clustered in two main groups; one mainly consisting of anelloviruses and one mainly of herpesviruses and papillomaviruses. Interestingly, taxonomically unrelated viruses were found to co-occur; BKV and Pegivirus A were associated with the anellovirus cluster, while human parvovirus and MCPyV were associated with papillomaviruses. The anellovirus cluster was associated primarily with leukaemias and mucosal samples, while the herpes and papillomavirus cluster was associated mainly with skin-associated and mucosal sample types.

Viruses with non-human hosts

Among the viral best BLASTnx hits for the contigs we identified hits to viruses from 25 viral families with non-human hosts, as well as unclassified viruses. The majority of these “non-human” viruses occurred ubiquitously across sample types (Fig S6) and detection of these seemed to be confined to the application of certain laboratory methods (Fig S7). These are considered in the Supplementary Results and in [23] (in press).

Evaluation of methods applied

Sequencing of total DNA or RNA, capture of retroviral DNA or mRNA, and mRNA enrichment showed few or no virus-positive samples. The remaining enrichment methods largely detected the same viral families, but not with the same frequency (Table 3, Fig S8-S9). Some of the viral findings were confirmed by more than one method (Fig S10). A comparison of the methods applied
in terms of number of samples positive, ability to retrieve high genome coverage or detect divergent viral sequences is presented in the Supplementary Results.

Discussion

In the present study we conducted a comprehensive virome investigation of 197 patient samples from 18 sample types of cancerous origin by applying a broad diversity of methods for enrichment of viral nucleic acids prior to sequencing. Targeting viruses with DNA and RNA genomes, double stranded, single stranded, and circular genomes, as well as proviruses, encapsidated and uncoated viral nucleic acids using sensitive enrichment methods (see Supplementary Discussion), we sought to fully cover the diversity of viruses present in the cancerous material. The resulting 710 distinct metagenomic datasets were analysed using a BLAST-based analysis approach and in-depth viral sequence analysis at both contig and read level. Our study provides central points of awareness concerning virome data analysis that need to be addressed before interpretation of the results. This includes viral artefacts, cross-mapping between closely related species/strains, and bleedover occurring during sequencing, as well as the presence of viral sequences in non-template controls (see Supplementary Material).

Most of the viruses identified in our study are commonly found in humans, and were almost exclusively DNA viruses (see Supplementary Discussion). Viral sequences were detected in a large percentage of the samples investigated and expectedly, skin-associated and mucosal samples showed higher proportions of virus-positive samples. Only few IARC-classified carcinogenic viruses were detected. These include the full genome of HPV16 identified in one of three vulvar cancer samples, confirming previous reports [39]. The full genome of HPV18 was detected in one of 10 bladder cancer urine samples. The evidence for a role of high-risk HPVs in the development of bladder cancer is currently inadequate [40],[41], and our study does not provide further support of high-risk HPVs playing a significant role. Evidence supports a causal role for HPV16 in a subset
of oropharyngeal cancers [1], whereas the prevalence of HPVs in oral cavity cancer is low [42]. The absence of high-risk HPVs is therefore not unexpected. Read mapping suggested presence of multiple HPV types in most HPV-positive samples. However, as was seen for BKV/JCV and HHV6A/6B, it cannot be ruled out that the detection of some types occur as a result of cross-mapping between closely related types. Viruses considered possibly carcinogenic and appearing in our samples include the polyomaviruses MCPyV, BKV, and JCV. These viruses are commonly carried asymptotically [43] and the findings could therefore represent normal flora.

A potential role for the ubiquitous anelloviruses in cancer is debated [44]. Multiple anelloviruses were often detected in the same sample, as previously reported in e.g. urine [12], however, no specific anellovirus types recurred consistently within cancer types. At the contig level, different species or strain can more readily be evaluated and distinguished (Table S11), however due to the read mapping patterns observed for some anelloviruses (see Supplementary Material) as well as possible cross-mapping, the diversity is possibly overestimated.

Parvovirus B19 was consistently detected in skin-associated samples. Seroprevalence is high in the general population, and the viral DNA can persist in multiple tissues, including skin [45],[46], although previous detection rates are lower that what was found here. Parvovirus B19 was not found in previously published skin and oral virome studies [5],[9], but these discrepancies could reflect differences in sample material and processing.

The effect of co-occurrence of viruses within a tissue is a relatively unexplored area. The co-occurrence of viral species and non-random distribution patterns found here reflect differences in viral tissue tropism, but other factors could play a role as well. Our study includes various habitats of the human body sampled from different individuals, providing a cross-body comparison of viral variation. Future studies of viral composition might reveal interactions of potential importance in health or disease between members of the virome.

With our study, several cancer types have been thoroughly investigated for viral nucleic acids. Cancer types investigated by us and not included in previous RNA-sequencing studies
[15],[16],[17],[47],[48] include basal cell carcinoma, testicular cancer, B-CLL, BCP-ALL, CML, T-ALL, vulvar cancer, and multiple myeloma cell lines. A limitation of our study is the low number of healthy control samples available, which hinders conclusions regarding viral presence in tumour versus normal flora. Although our sample size is not large, we consider the probability of uncovering yet undetected (known) viruses present in large proportions of these cancers low. HPV16 was detected in one of three vulvar cancer samples included, suggesting that our sample size is large enough to identify cancer-causing viruses of high prevalence. Nevertheless, low-frequency associations between known viruses and cancers might exist, and establishing causality in such cases is a complex process [49]. Other relevant approaches within cancer virus discovery include investigation of truly novel viruses with little or no similarity to known viruses, which are not detectable by the applied analysis methods. Moreover, changes in gene expression or DNA methylation may be directly induced by viral infections [50] and searching for such viral “footprints” could reveal new associations between previous viral infections and cancer.

Footnote page

Conflicts of interests
The authors declare no conflicts of interests.

Funding
This study was supported by the Innovation Fund Denmark (The GenomeDenmark platform, grant no. 019-2011-2), Danish National Research Foundation (grant no. DNRF94), and the Lundbeck Foundation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.
Meetings where parts of the results have previously been presented

Third International Conference on Clinical Metagenomics, Geneva, October 18-19 2018,

Acknowledgements

We thank Esben Nørgaard Flindt for valuable coordination of the GenomeDenmark platform. We thank Sarah Nathalie Vitcetz for technical assistance. We furthermore thank Zoltan Szallasi (Technical University of Denmark) for establishing contact to collaborators at National Institute of Oncology, Budapest, Hungary).

Author contributions

Conceived concept and designed the study: LPN, AJH, EW, LV, HF, SM, KRK, MA, JR, UB, SB.
Wrote the paper: SM.
Contributed significant input for editing of the manuscript: LV, AJH, MA, HF, KRK, JMGI, OL, LPN.
Scientific discussions for interpretation of data: SM, KRK, MA, LV, HF, AJH, LPN, TM.
Developed laboratory protocols: HF, LV, SM, KRK.
Performed laboratory experiments: SM, HF, LV, KRK, SRR, IBN, CP, ARI, DEAP, PVSO, RHJ.
Developed computational pipeline or provided supervision for this: JFN, JMGI, OL, TAH, AJH, TM, SM, SB, TSP.
Conducted initial bioinformatic analysis (pre-processing, assembly, BLAST, DIAMOND): JFN, JMGI, SM.
Performed further/additional/concluding analysis (data mining, investigation/confirmation of viral hits, visualisation): SM, MA, KRK, JFN, TM.
Mapping to viral genomes and creation of Circos plot: JARH.
Network analysis: CJB.
Conducted sample collection: ERDM, LGP, CvB, DHJ, RG, EH, IP, IV, ZB, KD, HJ, TS, PH, JLL, JR.
References


## Tables

### Table 1. Samples and datasets included in the study.

<table>
<thead>
<tr>
<th>Sample type</th>
<th>Sample material</th>
<th>Samples (n)</th>
<th>Total DNA</th>
<th>Total RNA</th>
<th>Virion enrichment</th>
<th>Capture</th>
<th>mDNA Datasets (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>DNA</td>
<td>RNA</td>
<td>DNA RNA enrich.</td>
<td>Virus DNA</td>
<td></td>
</tr>
<tr>
<td>Basal cell carcinoma (cutaneous)</td>
<td>Tumour biopsies</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>4</td>
<td>6</td>
<td>11</td>
</tr>
<tr>
<td>Mycosis fungoides (cutaneous)</td>
<td>Tumour biopsies</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>10</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>Melanoma (cutaneous)</td>
<td>Tumour biopsies</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>8</td>
<td>10</td>
<td>48</td>
</tr>
<tr>
<td>Oral cancer</td>
<td>Tumour biopsies</td>
<td>10</td>
<td>9</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>49</td>
</tr>
<tr>
<td>Oral healthy</td>
<td>Healthy tissue</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Vulvar cancer</td>
<td>Tumour biopsies</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td>Bladder cancer</td>
<td>Tumour biopsies</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>5</td>
<td>7</td>
<td>26</td>
</tr>
<tr>
<td>Bladder cancer urine</td>
<td>Urine</td>
<td>10</td>
<td>2</td>
<td>10</td>
<td></td>
<td></td>
<td>16</td>
</tr>
<tr>
<td>Colon cancer</td>
<td>Tumour biopsies</td>
<td>16</td>
<td>12</td>
<td>11</td>
<td>3</td>
<td>6</td>
<td>41</td>
</tr>
<tr>
<td>Colon healthy</td>
<td>Healthy tissue</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Breast cancer</td>
<td>Tumour biopsies</td>
<td>20</td>
<td>20</td>
<td>19</td>
<td>17</td>
<td>20</td>
<td>15</td>
</tr>
<tr>
<td>Testicular cancer</td>
<td>Tumour biopsies</td>
<td>20</td>
<td>5</td>
<td>20</td>
<td>20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AML</td>
<td>Bone marrow (sorted cells)</td>
<td>9</td>
<td>6</td>
<td>9</td>
<td>9</td>
<td>7</td>
<td>31</td>
</tr>
<tr>
<td>B-CLL</td>
<td>Blood/bone marrow (sorted cells)</td>
<td>9</td>
<td>8</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>BCP-ALL</td>
<td>Bone marrow</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td></td>
<td>24</td>
</tr>
<tr>
<td>CML</td>
<td>Bone marrow (sorted cells)</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>50</td>
</tr>
<tr>
<td>T-ALL</td>
<td>Bone marrow (non-sorted/sorted cells)</td>
<td>11</td>
<td>9</td>
<td>11</td>
<td>11</td>
<td>9</td>
<td>40</td>
</tr>
<tr>
<td>DLBCL</td>
<td>Cell lines</td>
<td>5</td>
<td>5</td>
<td></td>
<td>3</td>
<td>3</td>
<td>11</td>
</tr>
<tr>
<td>Lymphoblastic lymphoma</td>
<td>Cell lines</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Multiple myeloma</td>
<td>Cell lines</td>
<td>6</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Colon cancer blood</td>
<td>Blood</td>
<td>8</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Colon cancer ascites</td>
<td>Ascites</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Breast cancer ascites</td>
<td>Ascites</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Ovarian cancer ascites</td>
<td>Ascites</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>Pancreatic cancer ascites</td>
<td>Ascites</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>NTC</td>
<td></td>
<td>19</td>
<td>18</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td><strong>Total (without NTC)</strong></td>
<td></td>
<td>197</td>
<td>107</td>
<td>72</td>
<td>143</td>
<td>146</td>
<td>114</td>
</tr>
</tbody>
</table>

NTC: non-template control.
Table 2. Virus-positive samples from the read mapping analysis.

<table>
<thead>
<tr>
<th>Sample type</th>
<th>Samples (n)</th>
<th>Papillomaviridae</th>
<th>Polyomaviridae</th>
<th>Herpesviridae</th>
<th>Paroviridae</th>
<th>Anelloviridae</th>
<th>Flaviviridae</th>
<th>Retroviridae</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basal cell carcinoma</td>
<td>11</td>
<td>8</td>
<td>3</td>
<td>5</td>
<td>10</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mycosis fungoides</td>
<td>11</td>
<td>7</td>
<td>6</td>
<td>8</td>
<td>9</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Melanoma</td>
<td>10</td>
<td>7</td>
<td>3</td>
<td>6</td>
<td>8</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oral cancer</td>
<td>10</td>
<td>7</td>
<td>4</td>
<td>9</td>
<td>2</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oral healthy</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vulvar cancer</td>
<td>3</td>
<td>2</td>
<td></td>
<td></td>
<td>1</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bladder cancer</td>
<td>7</td>
<td>2</td>
<td>1</td>
<td>6</td>
<td>2</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bladder cancer urine</td>
<td>10</td>
<td>2</td>
<td>5</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Colon cancer</td>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Breast cancer</td>
<td>20</td>
<td>3</td>
<td>6</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Testicular cancer</td>
<td>20</td>
<td></td>
<td>2</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AML</td>
<td>9</td>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>B-CLL</td>
<td>9</td>
<td>1</td>
<td></td>
<td>3</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BCP-ALL</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CML</td>
<td>10</td>
<td>1</td>
<td></td>
<td>3</td>
<td>1</td>
<td>7</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>T-ALL</td>
<td>11</td>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Colon cancer blood</td>
<td>8</td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Colon cancer ascites</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ovarian cancer ascites</td>
<td>5</td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pancreatic cancer ascites</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total no. of samples</td>
<td>43</td>
<td>30</td>
<td>55</td>
<td>38</td>
<td>38</td>
<td>5</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Total no. of sample types</td>
<td>13</td>
<td>9</td>
<td>15</td>
<td>12</td>
<td>15</td>
<td>4</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

The number of samples positive for a given viral family is shown for each sample type. Extended counts are shown in Table S8. Only confirmed viral hits are included.
Table 3. Datasets positive for a given viral family for the laboratory methods applied.

<table>
<thead>
<tr>
<th>Method</th>
<th>Vert. virus capt. DNA</th>
<th>Circular DNA</th>
<th>Virion DNA</th>
<th>Virion RNA</th>
<th>Retrovirus capt. DNA</th>
<th>Total DNA</th>
<th>Total RNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>All samples</td>
<td>Datasets (n)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Papillomaviridae</td>
<td>Polyomaviridae</td>
<td>Herpesviridae</td>
<td>Parvoviridae</td>
<td>Anelloviridae</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vert. virus capt. DNA</td>
<td>75</td>
<td>3</td>
<td>31</td>
<td>1</td>
<td>3</td>
<td>10</td>
<td>42</td>
</tr>
<tr>
<td>Circular DNA</td>
<td>114</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Virion DNA</td>
<td>143</td>
<td>6</td>
<td>21</td>
<td>4</td>
<td>24</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>Virion RNA</td>
<td>146</td>
<td>1</td>
<td>13</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Retrovirus capt. DNA</td>
<td>33</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Total DNA</td>
<td>107</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Total RNA</td>
<td>72</td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

The number of datasets positive based on contig BLASTnx (grey columns) and read mapping (white columns) are shown. The top part of the table shows the numbers for all datasets, the bottom part shows the number for datasets from samples processed with all four enrichment methods. Only the five most frequently detected families are shown and only confirmed viral hits are included.

Non-template controls are excluded. \(^a P=9.5*10^{-5}\) vs. circular DNA enrichment, \(^b P=5.1*10^{-7}\) vs. virion enrichment DNA (non significant at contig level, \(P=0.061\)), \(^c P=4.6*10^{-4}\) vs. circular DNA enrichment (non significant at contig level, \(P=0.052\)), \(^d P=0.019\) vs. circular DNA enrichment.
Figure legends

Fig 1. Laboratory methods and analysis pipeline.
Top: Schematic illustration of the laboratory methods used. Total DNA or RNA was sequenced, or samples were exposed to one of the indicated enrichment methods prior to sequencing. Bottom: Schematic illustration of the data analysis pipeline; *de novo* assembled contigs and human-depleted reads were analysed with BLASTn and/or BLASTx/DIAMOND. Human viral hits were investigated *in silico*, and the reads were mapped to a database of selected viral reference genomes.
*Applies to the majority of the datasets, see Methods.

Fig 2. Viruses detected from BLASTnx of contigs and read mapping.
Top panel: The number of contigs detected across cancer types (horizontal axis), indicated by colour (right legend). Only confirmed viral hits are included. Bottom panel: The fraction of viral reads in parts per million (ppm) detected across cancer types (horizontal axis), indicated by colour (right legend). Only confirmed viral hits are included. NTC: non-template control.

Fig 3. HPVs identified in skin and mucosal cancers.
Genome coverage (%) for the different HPV types found in samples of skin and mucosal cancers, indicated by colour (right legend) (the full dataset is shown in Fig S4). Only confirmed viral hits are included.

Fig 4. Species co-occurrence network.
Network inference between the viruses grouped at species level. Nodes represent viral species, with diameters proportional to the total number of occurrences of a species (ranging from 4-40) and coloured segments representing the proportions of sample types a virus occurred in. Green colour tones represent skin-associated sample types, red/pink colour tones represent mucosal, blue
<table>
<thead>
<tr>
<th>Sample</th>
<th>Pegivirus</th>
<th>Human herpesvirus 4</th>
<th>Human herpesvirus 6</th>
<th>Cutavirus</th>
<th>Human parvovirus B19</th>
<th>Merkel cell polyomavirus</th>
<th>Human polyomavirus (BK/JC)</th>
<th>Alphapapillomavirus</th>
<th>Betapapillomavirus</th>
<th>Gammapapillomavirus</th>
<th>Unclass. Papillomavirus</th>
<th>Anellovirus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basal cell carcinoma</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mycosis fungoides</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Melanoma</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oral cancer</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oral healthy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vulvar cancer</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bladder cancer urine</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bladder cancer</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Breast cancer</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Testicular cancer</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AML</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B−CLL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BCP−ALL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CML</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T−ALL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Colon cancer ascites</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Colon cancer blood</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ovarian cancer ascites</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pancreatic cancer ascites</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NTC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>