Infectious Agents and Cancer

BRIEF COMMUNICATIONS Open Access

Metagenomic analysis to identify novel infectious agents in systemic anaplastic large cell lymphoma

Parag Mahale[1†], Jason Nomburg[2†], Joo Y. Song[3], Mia Steinberg[4], Gabriel Starrett[5], Joseph Boland[4], Charles F. Lynch[6], Amy Chadburn[7], Paul G. Rubinstein[8], Brenda Y. Hernandez[9], Dennis D. Weisenburger[3], Susan Bullman[10] and Eric A. Engels[1*]

Abstract

Systemic anaplastic large cell lymphoma (ALCL) is a rare CD30-expressing T-cell non-Hodgkin lymphoma. Risk of systemic ALCL is highly increased among immunosuppressed individuals. Because risk of cancers associated with viruses is increased with immunosuppression, we conducted a metagenomic analysis of systemic ALCL to determine whether a known or novel pathogen is associated with this malignancy. Total RNA was extracted and sequenced from formalin-fixed paraffin-embedded tumor specimens from 19 systemic ALCL cases (including one case from an immunosuppressed individual with human immunodeficiency virus infection), 3 Epstein-Barr virus positive diffuse large B-cell lymphomas (DLBCLs) occurring in solid organ transplant recipients (positive controls), and 3 breast cancers (negative controls). We used a pipeline based on the Genome Analysis Toolkit (GATK)-PathSeq algorithm to subtract out human RNA reads and map the remaining RNA reads to microbes. No microbial association with ALCL was identified, but we found Epstein-Barr virus in the DLBCL positive controls and determined the breast cancers to be negative. In conclusion, we did not find a pathogen associated with systemic ALCL, but because we analyzed only one ALCL tumor from an immunosuppressed person, we cannot exclude the possibility that a pathogen is associated with some cases that arise in the setting of immunosuppression.

Keywords: Lymphoma, Viruses, Metagenomics, Immunosuppression

Introduction

Systemic anaplastic large cell lymphoma (ALCL) is a rare CD30-expressing T-cell non-Hodgkin lymphoma (NHL) that comprises approximately 2% of all NHLs in adults [1]. Risk of systemic ALCL is markedly increased among immunosuppressed people, such as those with human immunodeficiency virus (HIV) infection and solid organ transplant recipients [2]. This increased risk in immunosuppressed populations suggests a viral etiology, because risk is similarly increased in immunosuppressed individuals for other lymphomas that are caused by Epstein-Barr virus (EBV), such as diffuse large B-cell lymphoma (DLBCL), Burkitt lymphoma, and Hodgkin lymphoma, as well as for other virus-related cancers [3]. EBV might be a plausible candidate as a cause of ALCL, but most reported ALCL tumors are EBV-negative [4].

There is a general lack of evidence regarding a possible role for infection in the etiology of systemic ALCL. Sequencing tumor tissue samples can provide a wealth of information about the biology of cancer, and while most sequences are human in origin, this approach can also detect the presence of novel viral agents. This approach

*Correspondence: engelse@exchange.nih.gov
†Parag Mahale and Jason Nomburg contributed equally to this work
[1] Infections and Immunoepidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD, USA
Full list of author information is available at the end of the article

was used successfully, for instance, in the identification of Merkel cell polyomavirus as the etiologic agent of Merkel cell carcinoma [5]. Metagenomics is the study of the pooled genetic material (genomes) from a mixed community of organisms. When applied to a human tissue specimen, the term can refer to the characterization of any microbial sequences among the much larger number of human sequences.

In the present study, we conducted a metagenomic analysis of systemic ALCL tumor RNA to assess for the presence of a known or novel pathogen in such cases, to help determine whether an infection may be implicated in the etiology of systemic ALCL.

Methods

Detailed methods are described in the Additional file 1.

Given the rarity of systemic ALCL (frequently referred to below simply as ALCL, for brevity), we obtained formalin-fixed paraffin-embedded (FFPE) tumor tissue from cases archived in the tumor tissue repositories of the Hawaii and Iowa cancer registries (N = 29). In addition, we leveraged a prior linkage of the Iowa cancer registry to the US solid organ transplant registry to identify DLBCL tumors in their repository occurring in transplant recipients (N = 3) as controls likely to be EBV-positive [6]. We obtained breast cancer cases from the Iowa cancer registry repository (N = 5) as negative controls because breast cancer does not have an established viral etiology. FFPE tissue from two additional ALCL cases in HIV-infected persons from Cook County Hospital (Illinois) and Weill Cornell Medicine (New York) were subsequently identified and included.

Four-micron tumor sections on charged slides were obtained for all ALCL and DLBCL tumors for histopathology and testing for EBV-encoded small RNAs (EBERs). One case (ID #IA25) that was originally reported as ALCL in the Iowa Cancer Registry was found to be misclassified DLBCL based on our pathology review. We reclassified this case as DLBCL and retained it in the study as an EBV-negative case (it was determined to be EBER-negative).

RNA was extracted from FFPE specimens, and total RNA libraries were prepared using the KAPA RNA HyperPrep Kit (Roche) and sequenced on a HiSeq 2500 platform (Illumina) (Additional file 1: Table S1, Supplementary Methods). Due to fragmentation of the RNA, libraries could be successfully prepared for only 19 of the 31 ALCL cases (61%, including only one HIV-infected case), 4 DLBCL controls (3 that were EBER-positive, plus IA25 which was EBER-negative), and 3 breast cancer controls, and these specimens were included in the metagenomic analysis.

A stepwise approach (Additional file 1: Fig. S1) was used for metagenomic classification of RNA-seq reads [7]. First, the Genome Analysis Toolkit (GATK)-PathSeq algorithm was used to perform computational subtraction of human reads, followed by alignments of residual reads to microbial reference genomes using BWA-MEM [8, 9].

To address the possibility that a known or novel pathogen was present in the reads that remained unclassified following GATK-PathSeq, we used the Viral Identification and Discovery (virID) pipeline in two approaches [7]. First, in assembly mode, the reads were assembled de novo into longer sequences (contigs) using rnaSPADES [10]. Contigs were taxonomically assigned using MegaBLAST [11], which aligns contigs with the NCI nucleotide "nt" reference database, and DIAMOND [12], which translates each contig into amino acid sequences and searches them against the RefSeq protein database [7]. Reads that were not assembled into assigned contigs in assembly mode were then analyzed with virID in read-mode (Supplementary Methods).

As a second alternative approach, all GATK-PathSeq unassigned reads were further assessed using a "kmer enrichment" strategy to identify non-repetitive 20 base pair sequences (20mers) present in at least two ALCL samples, none of the control tumors, and at least once in the RefSeq viral database [7]. We hypothesized that if a novel virus is associated with ALCL, then it would share 20mers with ALCL cases and other viral sequences, but not the control tumor specimens.

Results and discussion

Demographic and immunophenotypic characteristics of cases and controls are provided in Additional file 1: Table S2. Briefly, a majority of ALCL cases were ≤ 40 years of age at diagnosis (N = 10; 53%), men (N = 10; 53%), White (N = 16; 84%), and diagnosed in 2010 or earlier (N = 13; 68%). All cases tested negative for B-cell markers (CD20 or PAX5), all expressed CD30, and most expressed ALK1 (N = 11; 58%) and T-cell markers (CD2 or CD3; N = 10; 53%). All ALCL cases tested negative for EBER, whereas the three DLBCL controls occurring in transplant recipients were EBER-positive.

Additional file 1 also provides information on the input RNA and the total number of RNA reads for each specimen. Across the samples, the vast majority (99.8%) of RNA reads were human in origin, as expected. Using "approach 1" in Additional file 1: Fig. S1, a median of 69.5% of the remaining reads were assigned to known microorganisms (Fig. 1). Most classified reads were bacterial, except for the three EBER-positive DLBCL controls and one ALCL case (ID# IA16) where a high proportion of viral reads were observed (Additional file 1: Fig. S2A).

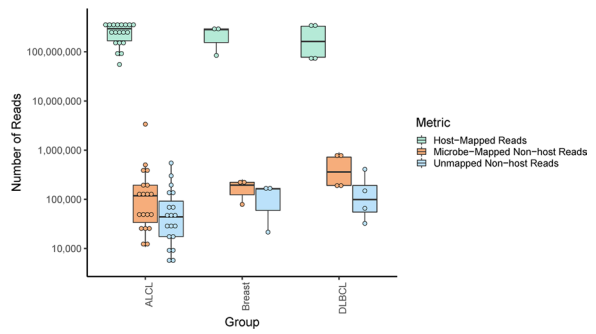

Fig. 1 GATK-PathSeq metrics. This figure presents the number of host (human) and non-host pathogen reads that were mapped by GATK-PathSeq and the non-host reads that remained unmapped. The number of reads was plotted as box plots on the y-axis and were divided into three groups: ALCL cases, DLBCL controls, and breast cancer controls. ALCL, anaplastic large cell lymphoma; DLBCL, diffuse large B-cell lymphoma

Hierarchical clustering of samples by their relative abundance of identified bacterial genera failed to classify the samples into specific tumor groups (Additional file 1: Fig. S3A, B).

GATK-PathSeq-based taxonomical classification did not identify any known viral genera preferentially associated with ALCL (Fig. 2). Of note, GATK-PathSeq mapped thousands of reads in the EBER-positive DLBCLs to the genus *Lymphocryptovirus*, which includes *human gammaherpesvirus 4* (EBV) and did not map reads to this genus in the EBER-negative DLBCL (ID# IA25) or breast cancer specimens. We identified a small number of EBV reads (n = 20) in ALCL case ID #IA16, which was EBER-negative, possibly due to contamination of the tumor specimen, presence of tumor-infiltrating lymphocytes containing EBV, or perhaps presence of a defective EBV genome within the tumor [13].

As shown in Fig. 2, there were many additional mapped viral genera among the RNA reads. However, these were distributed among both case and control groups, and/or the number of reads was very low (0–1 reads/million human reads). Furthermore, we observed that some virus genera with very few reads have plant hosts (e.g., *Polemovirus, Potexvirus, Ilarvirus*) or invertebrate hosts (e.g., *Seadornavirus* and *Tombusvirus*). Together, these observations suggest some degree of environmental contamination during tissue processing, storage, RNA extraction, library preparation, or sequencing [14–16], rather than a novel viral cause of ALCL.

In additional analyses of the RNA sequence data, the assembly and read-based modes of virID did not

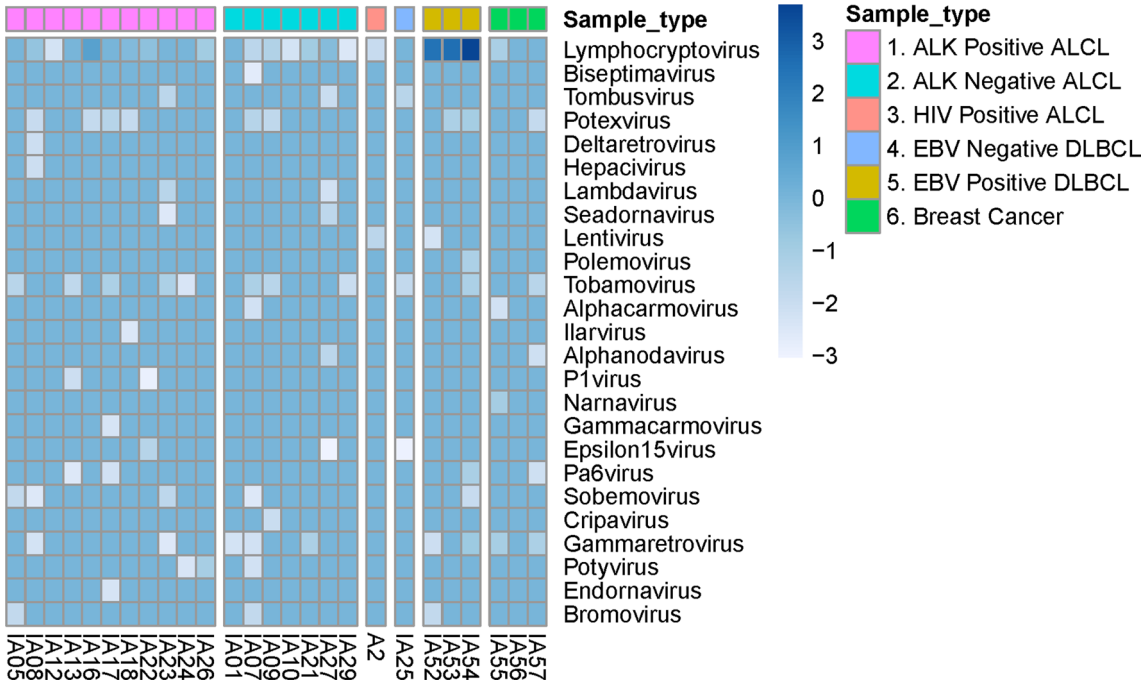

Fig. 2 GATK-PathSeq analysis of ALCL tumors. This figure shows the heat map GATK-PathSeq viral-mapped reads at the genus level. The units used are $\log_{10}$ reads per million human reads. Samples are grouped on the x-axis as ALK-positive ALCL, ALK-negative ALCL, HIV-positive ALCL, EBER-negative DLBCL, EBER-positive DLBCL, and breast cancer. Viral genera identified are listed on the y-axis. ALCL, anaplastic large cell lymphoma; ALK, anaplastic lymphoma kinase; DLBCL, diffuse large B-cell lymphoma; EBV, Epstein-Barr virus; HIV, human immunodeficiency virus

identify previously known or novel viruses in ALCL samples (Fig. 3, Additional file 1: Fig. S4A, B). Both approaches, however, again successfully identified viral reads belonging to the *Lymphocryptovirus* genus in at least one EBER-positive DLBCL specimen. To determine if any remaining but unassigned reads originated from repetitive human sequences, such as ribosomal sequences, we profiled the remaining reads with RepeatMasker. Of the reads that were not assigned to the human or microbial reference databases by GATK-PathSeq, a median of 19.7% remained unassigned following analysis with virID and RepeatMasker (Additional file 1: Fig. S2B), supporting that our analysis pipeline was able to identify and classify the substantial majority of microbial reads.

Finally, as an alternative approach ("approach 2" in Additional file 1: Fig. S1), we conducted a BLASTN search of reads containing enriched 20mers from ALCL samples and identified Hubei tombus-like virus 8 and Norway luteo-like virus 4. However, these viruses have invertebrate hosts and are possibly environmental contaminants (Additional file 1: Fig. S5A, B) [14].

We leveraged two population-based cancer registries to identify and obtain FFPE blocks from systemic ALCL cases, which enabled us to study this rare malignancy. An additional strength of our study was our detailed phenotyping of the tumors to confirm the ALCL diagnosis and the absence of EBV. Our reliance on archived FFPE blocks that were stored for several years resulted in highly fragmented RNA that likely reduced our

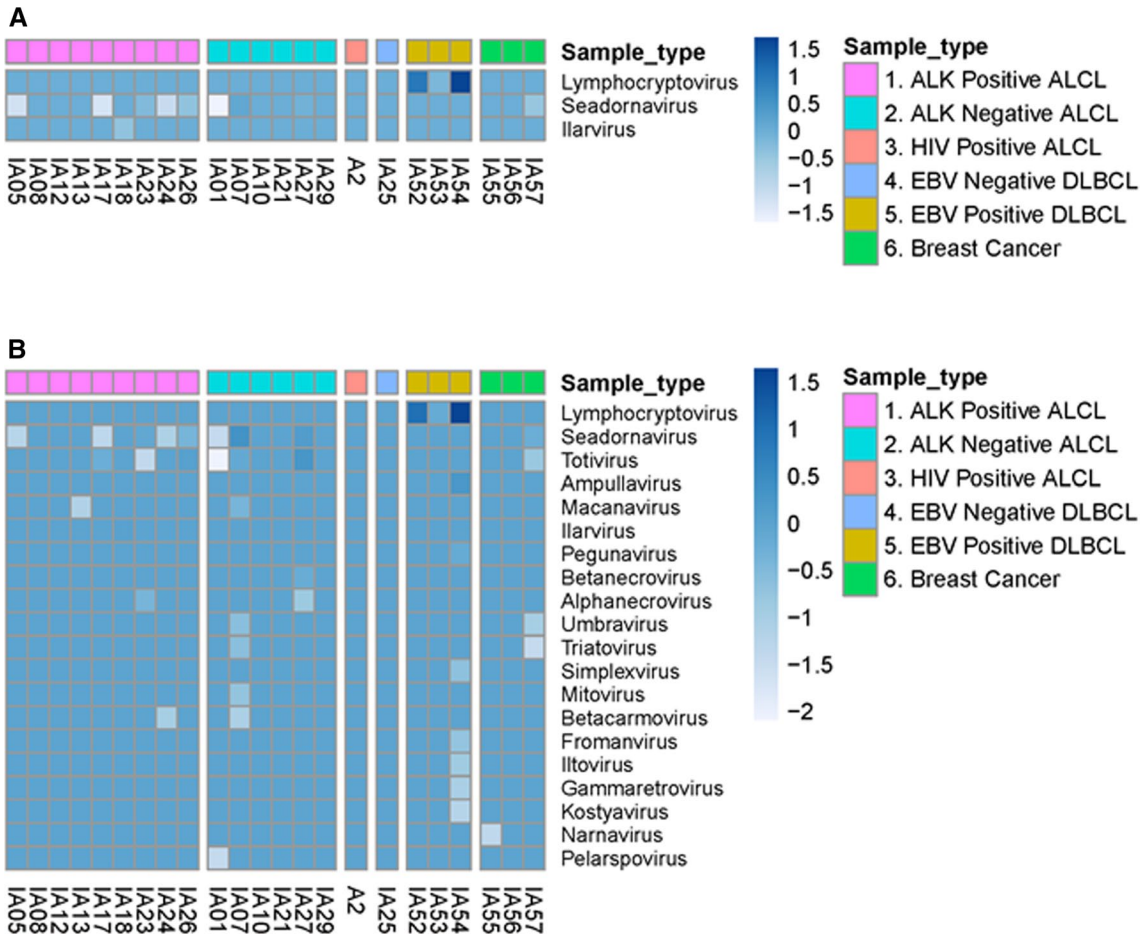

Fig. 3 Application of virID pipeline to identify viral sequences associated with ALCL. This figure highlights the findings of applying virID algorithm to classify unmapped reads following GATK-PathSeq using the virID assembly-based approach. **A** and **B** represent the taxonomical classification of reads into viral genera after subjecting the contigs to nucleotide (MegaBLAST) and translated amino acid (DIAMOND) searches against their respective reference databases. The units used are $\log_{10}$ reads per million human reads. Samples are grouped on the x-axis as ALK-positive ALCL, ALK-negative ALCL, HIV-positive ALCL, EBER-negative DLBCL, EBER-positive DLBCL, and breast cancer. Viral genera identified are listed on the y-axis. ALCL, anaplastic large cell lymphoma; ALK, anaplastic lymphoma kinase; DLBCL, diffuse large B-cell lymphoma; HIV, human immunodeficiency virus

sensitivity, and this may also have contributed to the number of contaminating reads. Nonetheless, we demonstrated the sensitivity of our sequencing and bioinformatics approach by successfully identifying EBV reads in all three EBER-positive control DLBCL specimens, which would suggest high overall sensitivity. We did not identify EBV or other novel viruses in breast cancer specimens that we used as negative controls.

Because we analyzed only one ALCL tumor from an HIV-infected person, we cannot exclude the possibility that a pathogen is associated with some ALCL cases that arise among immunosuppressed people. A small number of EBV-positive ALCL cases have been described in people living with HIV and solid organ transplant recipients [17–19]. ALCL can resemble other T-cell lymphomas, and differential diagnosis among these subtypes is challenging. In particular, ALK1-negative ALCLs must be distinguished from extranodal NK/T-cell lymphomas, which are strongly associated with EBV infection. ALCLs exhibit characteristic "hallmark" cells, and tumor cells show diffuse and strong CD30 expression. In contrast, CD30 expression in extranodal NK/T-cell lymphoma is usually more variable. ALCLs can be positive for CD4 (which is not typically seen with extranodal NK/T-cell lymphoma), and many null-type ALCLs lack CD3 (whereas cytoplasmic CD3 expression is retained in extranodal NK/T-cell lymphoma).

In conclusion, in the first metagenomic analysis of systemic ALCLs, we did not find a pathogen that was associated with this rare cancer. Our study had a modest sample size, and further studies are required to characterize the metagenome of systemic ALCLs, especially cases arising in immunosuppressed people.

Abbreviations

ALCL: Anaplastic large cell lymphoma; DLBCL: Diffuse large B-cell lymphoma; EBV: Epstein-Barr virus; EBER: EBV-encoded small RNAs; FFPE: Formalin-fixed paraffin-embedded; GATK: Genome Analysis Toolkit; HIV: Human immunodeficiency virus; NHL: Non-Hodgkin lymphoma.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13027-021-00404-0.

Additional file 1: Figure S1. Schematic of the computational approach used in the metagenomic analysis of ALCL. **Figure S2**. Additional GATK-PathSeq analysis of tumor specimens. **Figure S3**. Detailed taxonomic classification of GATK-PathSeq assigned non-human reads. **Figure S4**. Use of read-based approach to identify viral sequences associated with ALCL. **Figure S5**. The kmer enrichment approach to identify pathogen reads from unmapped GATK-PathSeq non-human reads. **Table S1**. Quality control results for cases included in the analysis. **Table S2**. Demographic and immunophenotypic features of cases and controls. **Table S3** Immunohistochemistry and in situ hybridization panel performed on ALCL and DLBCL tumor specimens.

Acknowledgements

The authors acknowledge the research contributions of the Cancer Genomics Research Laboratory for their expertise, execution, and support of this research in the areas of project planning, wet laboratory processing of specimens, and bioinformatics analysis of generated data. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government.

Authors' contributions

The study was designed by PM and EAE. Study samples were provided by CFL, AC, PGR, and BYH. Pathologic review was performed by JYS, AC, and DDW. RNA extraction and sequencing were performed by MS, GS, and JB. Sequence analyses were conducted by JN, MS, GS, and SB. PM, JN, and EAE wrote the manuscript. All other authors reviewed the results and provided critical editorial suggestions for the manuscript. All authors read and approved the final manuscript.

Funding

Open Access funding provided by the National Institutes of Health (NIH). Funding was provided by the Intramural Research Program of the National Cancer Institute.

Availability of data and materials

RNA sequencing data are available in the database of Genotypes and Phenotypes (dbGaP) under accession number phs002064.v1.p1.

Declarations

Ethics approval and consent to participate

This study was approved by the institutional research boards of the University of Hawaii and the University of Iowa. It was considered exempt from human subjects review by the National Institutes of Health. Patient consent was not required because the study utilized residual tumor specimens obtained for clinical purposes and the genomic data were not analyzed in association with personal identifiers.

Consent for publication

The authors provide consent for publication.

Competing interests

The authors have no competing interests.

Author details

[1]Infections and Immunoepidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD, USA. [2]Department of Medical Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, USA. [3]Department of Pathology, City of Hope National Medical Center, Duarte, CA, USA. [4]Cancer Genomics Research Laboratory, National Cancer Institute, Rockville, MD, USA. [5]Laboratory of Cellular Oncology, Center for Cancer Research, National Cancer Institute, Bethesda, MD, USA. [6]Department of Epidemiology, The University of Iowa College of Public Health, Iowa City, Iowa, USA. [7]Pathology and Laboratory Medicine, Weill Cornell Medical College, New York, NY, USA. [8]Stroger Hospital of Cook County, Ruth M. Rothstein Core Center, Rush University Medical Center, Chicago, IL, USA. [9]University of Hawaii Cancer Center, Honolulu, HI, USA. [10]Human Biology Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA.

Received: 24 August 2021 Accepted: 3 November 2021
Published online: 14 November 2021

References

1. Morton LM, Wang SS, Devesa SS, Hartge P, Weisenburger DD, Linet MS. Lymphoma incidence patterns by WHO subtype in the United States, 1992–2001. Blood. 2006;107(1):265–76.
2. Mahale P, Weisenburger DD, Kahn AR, Gonsalves L, Pawlish K, Koch L, et al. Anaplastic large cell lymphoma in human immunodeficiency

virus-infected people and solid organ transplant recipients. Br J Haematol. 2020;192:514–21.
3. Grulich AE, van Leeuwen MT, Falster MO, Vajdic CM. Incidence of cancers in people with HIV/AIDS compared with immunosuppressed transplant recipients: a meta-analysis. Lancet. 2007;370(9581):59–67.
4. Herling M, Rassidakis GZ, Jones D, Schmitt-Graeff A, Sarris AH, Medeiros LJ. Absence of Epstein-Barr virus in anaplastic large cell lymphoma: a study of 64 cases classified according to World Health Organization criteria. Hum Pathol. 2004;35(4):455–9.
5. Feng H, Shuda M, Chang Y, Moore PS. Clonal integration of a polyomavirus in human Merkel cell carcinoma. Science. 2008;319(5866):1096–100.
6. Engels EA, Pfeiffer RM, Fraumeni JF Jr, Kasiske BL, Israni AK, Snyder JJ, et al. Spectrum of cancer risk among US solid organ transplant recipients. JAMA. 2011;306(17):1891–901.
7. Nomburg J, Bullman S, Chung SS, Togami K, Walker MA, Griffin GK, et al. Comprehensive metagenomic analysis of blastic plasmacytoid dendritic cell neoplasm. Blood Adv. 2020;4(6):1006–11.
8. Walker MA, Pedamallu CS, Ojesina AI, Bullman S, Sharpe T, Whelan CW, et al. GATK PathSeq: a customizable computational tool for the discovery and identification of microbial sequences in libraries from eukaryotic hosts. Bioinformatics (Oxford, England). 2018;34(24):4287–9.
9. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. https://arxiv.org/abs/1303.3997. Accessed April 14, 2020. https://arxiv.org/abs/1303.3997.
10. Bushmanova E, Antipov D, Lapidus A, Prjibelski AD. rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. GigaScience. 2019;8(9).
11. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinform. 2009;10:421.
12. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12(1):59–60.
13. Gan YJ, Razzouk BI, Su T, Sixbey JW. A defective, rearranged Epstein-Barr virus genome in EBER-negative and EBER-positive Hodgkin's disease. Am J Pathol. 2002;160(3):781–6.
14. Shi M, Lin XD, Tian JH, Chen LJ, Chen X, Li CX, et al. Redefining the invertebrate RNA virosphere. Nature. 2016;540(7634):539–43.
15. Tang K-W, Larsson E. Tumour virology in the era of high-throughput genomics. Philos Trans R Soc Lond Ser B Biol Sci. 2017;372(1732).
16. Asplund M, Kjartansdóttir KR, Mollerup S, Vinner L, Fridholm H, Herrera JAR, et al. Contaminating viral sequences in high-throughput sequencing viromics: a linkage study of 700 sequencing libraries. Clin Microbiol Infect. 2019;25(10):1277–85.
17. Perez K, Castillo J, Dezube BJ, Pantanowitz L. Human Immunodeficiency Virus-associated anaplastic large cell lymphoma. Leukemia Lymphoma. 2010;51(3):430–8.
18. Herreman A, Dierickx D, Morscio J, Camps J, Bittoun E, Verhoef G, et al. Clinicopathological characteristics of posttransplant lymphoproliferative disorders of T-cell origin: single-center series of nine cases and meta-analysis of 147 reported cases. Leukemia Lymphoma. 2013;54(10):2190–9.
19. Pitman SD, Rowsell EH, Cao JD, Huang Q, Wang J. Anaplastic large cell lymphoma associated with Epstein-Barr virus following cardiac transplant. The Am J Surg Pathol. 2004;28(3):410–5.

Publisher's Note