JCI insight

Proteomic approach to discover human cancer viruses from formalin-fixed tissues

Tuna Toptan, … , Yuan Chang, Patrick S. Moore

JCI Insight. 2020. https://doi.org/10.1172/jci.insight.143003.

Resource and Technical Advance In-Press Preview Virology

The challenge of discovering a completely new human tumor virus of unknown phylogeny or sequence depends on detecting viral molecules and differentiating them from host molecules in the virus-associated neoplasm. We developed differential peptide subtraction (DPS) using differential mass-spectrometry (dMS) followed by targeted analysis to facilitate this discovery. We validated this approach by analyzing Merkel cell carcinoma (MCC), an aggressive human neoplasm, in which ~80% of cases are caused by the human Merkel cell polyomavirus (MCV). Approximately 20% of MCC have a high mutational burden and are negative for MCV, but are microscopically indistinguishable from virus positive cases. Using 23 (12 MCV positive, 11 MCV negative) formalin-fixed MCC, DPS identified both viral and human biomarkers (MCV Large T antigen, CDKN2AIP, SERPINB5 and TRIM29) that discriminates MCV positive and negative MCC. Statistical analysis of 498,131 dMS features not matching the human proteome by DPS revealed 562 (0.11%) to be up-regulated in virus-infected samples. Remarkably, four (20%) of the top 20 candidate MS spectra originated from MCV T oncoprotein peptides and confirmed by reverse translation degenerate oligonucleotide sequencing. DPS is a robust proteomic approach to identify novel viral sequences in infectious tumors when nucleic acid-based methods are not feasible.

Proteomic approach to discover human cancer viruses from formalin-fixed tissues

Tuna Toptan[1,2], Pamela S. Cantrell[3], Xuemei Zeng[3], Yang Liu[3], Mai Sun[3], Nathan A. Yates[3,4*], Yuan Chang[1,*], Patrick S. Moore[1,*]

[1] Hillman Cancer Center, University of Pittsburgh, Pittsburgh, PA, 15232, USA

[2] Institute of Medical Virology, University Hospital Frankfurt, Goethe University, Frankfurt am Main, 60590, Germany

[3] Biomedical Mass Spectrometry Center, University of Pittsburgh, Pittsburgh, PA, 15261, USA

[4] Department of Cell Biology, University of Pittsburgh, Pittsburgh, PA, 15261, USA

* Equal contribution

The authors have declared that no conflict of interest exists.

ABSTRACT

The challenge of discovering a completely new human tumor virus of unknown phylogeny or sequence depends on detecting viral molecules and differentiating them from host molecules in the virus-associated neoplasm. We developed differential peptide subtraction (DPS) using differential mass-spectrometry (dMS) followed by targeted analysis to facilitate this discovery. We validated this approach by analyzing Merkel cell carcinoma (MCC), an aggressive human neoplasm, in which ~80% of cases are caused by the human Merkel cell polyomavirus (MCV). Approximately 20% of MCC have a high mutational burden and are negative for MCV, but are microscopically indistinguishable from virus positive cases. Using 23 (12 MCV positive, 11 MCV negative) formalin-fixed MCC, DPS identified both viral and human biomarkers (MCV Large T antigen, CDKN2AIP, SERPINB5 and TRIM29) that discriminates MCV positive and negative MCC. Statistical analysis of 498,131 dMS features not matching the human proteome by DPS revealed 562 (0.11%) to be up-regulated in virus-infected samples. Remarkably, four (20%) of the top 20 candidate MS spectra originated from MCV T oncoprotein peptides and confirmed by reverse translation degenerate oligonucleotide sequencing. DPS is a robust proteomic approach to identify novel viral sequences in infectious tumors when nucleic acid-based methods are not feasible.

INTRODUCTION

Seven human viruses are responsible for approximately 15% of the tumor burden world-wide. This phylogenetically heterogeneous group of viruses differ extensively in their genome sizes, nucleic acid composition, and replication mechanisms (1). Likewise, the discovery processes for each of these seven tumor viruses has varied and evolved closely with technological advances, particularly in molecular biology. Epstein-Barr virus (EBV or human herpesvirus (HHV4) (2)), a large double stranded DNA herpesvirus, was first identified in 1964 based on classic microbiology detection practices in cell culture and electron microscopy. Hepatitis B virus (HBV), unculturable at the time, was found by serologic screening in 1965 (3). The discovery of human T-lymphotropic virus-1 (HTLV-1), a retrovirus, was facilitated by reverse transcriptase assays in 1980 (4), and, although many strains of human papillomaviruses (HPV) were already identified by 1983, cervical cancer-associated HPV strains were only identified through strain-specific DNA Southern hybridization studies (5). Hepatitis C virus (HCV), a flavivirus, was found by cDNA library screening in 1989 (6).

Molecular subtractive techniques have been most recently used to determine the infectious etiologies of Kaposi sarcoma (KS) and Merkel cell carcinoma (MCC). In 1993, fragments of the Kaposi sarcoma herpesvirus (KSHV/HHV8) genome were cloned using representational difference analysis (RDA), a DNA-based subtractive process which can isolate foreign nucleic acids from the human genome (7, 8). In 2008, Merkel cell polyomavirus (MCV) transcripts were found by digital transcriptome subtraction (DTS), an *in silico* RNA subtractive process taking advantage of a timely expansion in sequencing capabilities, databases and search engines (9, 10).

Virus-associated cancers are biological accidents, detrimental to both the host and the viral pathogen (11). The cancer virus is generally not actively replicating (latency or pseudo-

latency) in cancerous cells, which would otherwise tend to kill the host cell. However, latent viral transcript levels tend to be reduced relative to cellular or lytic viral transcript levels (12). Latent viral proteins, on the other hand, can have exceptional stability (13), be expressed by non-canonical translation (14, 15), and can circumvent cellular protein degradation mechanisms (13). This is thought to be a viral strategy to reduce immunoproteasomal peptide processing to escape host immune responses against latent viral proteins (16). Based on these biologic features of tumor viruses, we pursued a protein-based detection method that may be useful for tumors in which RNA is unavailable or in which viral transcript levels are too low to be routinely detected. A protein-based virus discovery method using cross-reactive antibodies to viral proteins has been described for polyomaviruses (pan-polyomavirus immunohistochemistry test, P-PIT) (17, 18). Since P-PIT depends on conserved epitope(s) within a class of known viruses, it cannot identify unique nucleic acid or peptide sequences from a new agent.

We sought an unbiased approach for deep peptide sequencing to differentiate human from foreign peptides belonging to novel viruses that can make use of archival pathology samples. To achieve this, we developed a methodology called differential peptide subtraction (DPS) using label-free differential mass-spectrometry (dMS) to quantify relative peptide abundance between complex samples (19-21). The advantages of DPS are that it is able to interrogate protein abundance, can identify novel peptides, makes use of formalin-fixed, paraffin-embedded (FFPE) tissues, and, if no pathogen is found, can reveal unique cellular protein biomarkers that may improve diagnosis and prognosis of a target disease.

RESULTS

Unbiased DPS was performed on polyomavirus-positive and -negative MCC. MCC is a highly aggressive human skin cancer, 80% of which is etiologically associated with MCV (9). Virus positive MCCs express viral small T (sT) and large T (LT) antigen oncoproteins and have a low mutation burden (22). In this subset of MCCs, MCV genome clonally integrates into the host chromosome and acquires mutations or deletions resulting in the translation of C-terminally truncated LT proteins which vary in size from tumor to tumor. In contrast, virus-negative MCC, although microscopically indistinguishable from virus-positive MCC, carry high mutation burdens and driver somatic mutations that phenocopy MCV infection (23). We used MCV negative (n=11) and positive (n=12) MCC FFPE tissues, and processed them in a blinded fashion to determine if DPS can distinguish the presence of a tumor virus *de novo* in human tissues without prior knowledge of the virus identity or sequence. The dMS workflow that was used for sample processing and data analysis is illustrated in the graphical abstract.

Polyomavirus status was initially determined by MCV LT antigen immunohistochemical staining (**Supplemental Table 1**). Proteins were extracted from FFPE tissues and digested using filter-aided sample preparation (FASP)-based tryptic digestion and analyzed by nano-flow liquid chromatography tandem mass spectrometry (nLC-MS/MS) (**Figure 1A**). High-resolution full-scan (MS1) mass spectra and low-resolution tandem (MS2) mass spectra were recorded on a hybrid ORBItrap Velos mass spectrometer (**Figure 1A**). Four types of experimental samples were included in the experimental design: eleven MCV negative samples, twelve MCV positive samples, nine sample processing replicates, and one instrument control sample (**Figure 1A**, **Supplemental Figure 1**). A data set of 498,131 high-resolution MS1 features (**Supplemental Table 2**) was extracted from the raw mass spectral data using the MaxQuant (v1.6.0.1) proteomic software package (24, 25). Subsequently, all MS1 features that could be identified by searching the MS2 spectra

against a human Uniprot protein database (downloaded in February 2013 with 87,662 entries) were removed. The log2 transformed intensities of the unidentified proteomic features were analyzed with a Student's t test to select features that exhibit significant differences in relative abundance between MCV-positive and MCV-negative tumor samples. Filtering for spectral features with a p-value < 0.01 and at least a 10-fold higher intensity in MCV-positive samples compared to MCV-negative samples returned 562 features. Targeted nLC-MS/MS analysis was used to collect MS2 spectra for the 20 most significant features ranked by ascending p-value. Manual *de novo* sequencing identified amino acid sequence tags greater than five amino acids long for 11 of 20 selected features (**Table 1**). A Blast search against UniProtKB revealed that four of these peptides matched to the MCV T antigen protein sequence (**Table 1, Figure 1B**). MCV small T (sT) and Large T (LT) antigens are derived from differentially-spliced transcripts and share a 78-amino acid (aa) N-terminus, nucleotide 196-429, Frame 1). LT splices after the first exon into a C-terminal exon (738 aa, nucleotide 861-3080, Frame 3), whereas sT-transcript reads through the initial splice site **(Figure 1B)**. The localization of four peptides identified by targeted nLC-MS/MS analysis are shown in Figure 1B. The two peptides on the left (green-ID#14 and orange-ID#4) are common to sT and LT, whereas the other two (purple-ID#15 and blue-ID#1) correspond to LT. There is a partial match to sT (DE) for the third peptide (purple-ID#15) which spans the splice junction between exon 1 and 2 of LT. The relative abundance of the identified MCV peptides in the infected vs. control samples (**Figure 1C**) show that a human tumor virus in tumor tissues can generate sufficient protein to be identified *de novo* from tumor tissue. Thus, the comparison of proteomic profiles from infected and control tissues allows identification of new proteins without *a priori* knowledge of the protein sequence.

In the case of a novel virus, dMS-identified peptides will not have a match in the databases, nevertheless this information can facilitate the recovery of the viral genome sequence. To

this end, we sought to trace the non-human dMS identified peptides back to their genetic origins by next generation sequencing (NGS) with cDNA libraries generated using degenerate oligos based on the identified peptide sequences (**Supplemental Table 3**). In designing degenerate primers, we aimed to avoid primer sequences with 6-fold codon nucleotide variants (L, S and R) and to maximize the number of 2 fold codon variants (D, E, Y, N, K) thereby maintaining moderate binding-specificity while reducing oligo degeneracy. In line with this, peptide areas with "X" residues (Table 1) representing either an L or an I (3-fold degeneracy sites) codons which are indistinguishable by MS were excluded. In addition to non-human matches (ID#: 1,4, and 15, **Table 1**), we included peptide ID# 3 which was only a partial match to human in a Blast search. Based on the in silico reverse translation, forward and reverse primers were designed for a total of four peptides (**Supplemental Table 3**). First, we tested the binding efficiency of these degenerate primers (DP) by a low-cycle RT-PCR (**Figure 2A**). Four different sets of combinations of forward and reverse DP, and cDNA template from a MCV positive MCC sample were used for the PCR reaction (**Figure 2A**). Combinations of Forward (F)4, Reverse (R)1 and F4, R15 primers resulted in 400 and 200 bp PCR products respectively, which were confirmed to be derived from MCV by sequencing (**Figure 2B**).

For the library generation to perform NGS analysis, degenerate oligos were fused to SMART adaptor sequence (SMART-deg, 25 nt) (**Supplemental Table 4**) and used for cDNA synthesis as described previously (26) (**Figure 2C**). To demonstrate the principle and efficiency of this procedure, mixtures of SMART-degenerate oligos or a modified oligo(dT) SMART primer were used to facilitate reverse transcription from viral or viral and host RNA, respectively. Due to high degeneracy of these primers we sought to increase their specificity and designed another set of primers by addition of a number of locked-nucleic acid (LNA) modifications for indicated bases (**Supplemental Table 4**). Using these new LNA modified primers, we generated MCC^{deg} $MCC^{LNA\text{-}deg}$, and (MCC^{polyA}) SMART-cDNAs which were then

processed into three Nextera Flex libraries and subsequently sequenced using NextSeq 500 (**Figure 2D**). 58-68 million reads per sample were obtained which were processed and mapped to a combined reference index from GRCh38 and MCV (JF813003) annotations (**Table 2**). Normalization procedures to account for different sequencing depths amongst the three libraries include conversion of data to transcripts-per-million (TPM) read-outs and trimmed mean of M values (TMM). We detected 7.3 and 2.6 times more MCV reads in degenerate oligo primed RNAseq samples compared to polyA-based sequencing reads. In addition, reads from MCC^{deg} and $MCC^{LNA\text{-}deg}$ largely mapped upstream of the DP binding sites within the T antigen region. Hence, this strategy can facilitate the identification of a viral genome sequences even in cases where the dMS-peptides do not match to previously identified pathogens.

Label-free dMS method not only identified differentially expressed viral peptides within a complex mixture, but also proteins which can serve as prognostic biomarkers. A total of 17,921 unique human peptides from 2,832 corresponding protein groups were quantified and the peptide intensity values were log2 transformed (**Supplemental Table 5**). A Student's t test was used for statistical comparison between MCV-positive and –negative peptide intensity values. Significant proteins were selected if more than half of the identified peptides from a protein were significant (p value < 0.05) and single peptide identifications were excluded from the analysis. A total of 38 proteins showed significantly increased abundance whereas eight proteins were decreased in abundance in MCV-positive samples. The list of identified peptides for these proteins are included in **Supplemental Table 6**.

To validate differentially expressed human peptides as potential biomarkers, five MCV positive and four MCV negative MCC tissue cores together with control tissues were used to generate a tissue microarray and were analyzed for the expression of CDKN2AIP, SERPINB5 and TRIM29 by immunohistochemistry (**Figure 3)**. Consistent with dMS results,

193 we found loss of TRIM29 and SERPINB5 expression, and higher levels of CDKN2AIP
194 expression in all MCV positive MCC cases (**Figure 3A**). These results suggest a role of
195 MCV T antigens in the regulation of SERPINB5 and TRIM29 expression (**Figure 3B**).
196

DISCUSSION

In this study, we provide a nanoLC-MS/MS-based protocol to compare tissues and identify differentially expressed peptides and potential prognostic markers. This is a peptide/proteome subtraction process that is analogous to the mRNA/transcriptome subtraction (DTS) originally used to discover MCV (9). Importantly, the high DPS de novo identification rate for MCV peptides in the context of the entire human tumor tissue proteome shows that this approach is promising. We anticipate that it can supplement RNA-based analyses of suspected infectious cancers, especially for tumors in which it is difficult to obtain sufficient RNA for sequencing.

The top 20 unsupervised candidate MS feature sequences (after differential and database subtraction that were present in MCV-positive but not MCV-negative samples) were manually determined, a laborious process. These 20 peptide sequences were then aligned to the human proteome by BLAST-P, which revealed 4 of these 20 peptides to be of MCV origin. These four peptides map to the N-terminus of the MCV T antigen oncoprotein complex, including peptides common to sT and LT and to the beginning of the second exon in LT, that are common to the coding regions of the truncated LT proteins found in all the MCV-positive MCC tumors (**Figure 1B**).

Modern virus discovery only requires a discovery of a single unique nucleotide sequence to recover the entire viral sequence by gene walking. We show that starting from three unique peptides, NGS of degenerate cDNA from MCV positive MCC tumor library recovers unique viral nucleic acid sequences that can allow full viral characterization. Although this approach proved to be more efficient than poly-A NGS, LNA modifications to the oligos used for cDNA generation did not seem to improve the outcome. We anticipate that sequentially performing these steps (first DPS on formalin-fixed tumor tissues followed by degenerate NGS of

candidate peptide coding sequences using well-accessioned tumor tissue RNA libraries) is a viable strategy to find and characterize human tumor viruses, particularly in rare tumors.

DPS relies on comparison of a viral cancer proteome to a matched control non-viral tumor proteome. Other known paired viral/non-viral tumors that could be similarly tested include head-and-neck carcinoma, nasopharyngeal carcinoma, Burkitt lymphoma, and hepatocellular carcinoma (1). In our study spectral features with a p-value < 0.01 and at least a 10-fold higher intensity in MCV-positive tumors returned 562 features. Targeted nLC-MS/MS analysis was used to collect MS2 spectra for the 20 most significant features ranked by ascending p-value which enabled the identification of candidate viral peptides. Effective ranking and prioritization is important because *de novo* sequencing remains a largely tedious and slow manual process. The entire protocol from the processing of blinded specimens to the unbiased identification of the viral protein consumes less than three weeks of laboratory time.

For some of the cancer types, however, a well-defined control group might not be available. In such cases, statistical analysis for hierarchical clustering of the samples might be useful. To specifically address this potential problem, we used un-supervised hierarchical clustering to investigate the possibility of using proteomic profiles to accurately classify MCV samples into two groups, viral positive and viral negative. The best classification result was obtained using the proteomic profiles for proteins associated with virus-related biological processes. A total of 19 out of 22 samples were correctly classified (86% accuracy, **Supplemental Figure 2**). MCV features remained significantly different between the two cluster groups despite the drop in their significance ranking, supporting the potential of applying un-supervised clustering for classification of samples with unknown viral status.

An alternative approach in the absence of matched negative control tissues is the generation of a reference database comprising human MS/MS peptide features. This reference database could then be used for DPS in silico subtraction of universal “human” peptides from tumor MS/MS profiles, leaving candidate “non-human” peptide sequences. Such a MS/MS database (the proteome equivalent of the nucleotide RefSeq database) does not currently exist. Such a database would also be highly dependent on machine and sample characteristics, as well as biological characteristics (e.g., single nucleotide polymorphisms, post-transcriptional and post-translational modifications) that would make universal comparisons difficult. Following subtraction, degenerate NGS using the candidate non-human tandem MS spectra to design oligonucleotides could be used to search for viral sequences. This strategy would not only circumvent the need for well-matched histological tissue controls but also reduce the cost, time and manual labor needed in evaluation steps. As with nucleotide DTS (10), an in silico DPS analysis may miss a viral pathogen if commensal or endogenous virus peptide features are mistakenly assigned as “human” in the comparison database.

Even when no new virus is found, DPS has utility for identifying human protein biomarkers. We identified 38 human proteins significantly increased and eight proteins decreased in MCV-positive vs. MCV-negative MCC samples, including SERPINB5, a reported tumor suppressor also known as mammary serine protease inhibitor (MASPIN) (27), and TRIM29, a ubiquitin E3 ligase that may act as a scaffold protein in the DNA damage response (28). Loss of TRIM29 expression promotes invasion of skin squamous cell carcinoma cells by altering distribution of keratins (29). Loss of expression of these two proteins might contribute to a more aggressive disease course for MCV-negative compared to MCV-positive MCC; however, larger cohort studies are needed to confirm these initial findings. These and other differentially expressed proteins can be readily examined as potential

prognostic biomarkers for MCV-positive and MCV-negative MCC tumors or as biomarkers to differentiate MCV-negative MCC from other small round cell neuroectodermal cancers.

DPS also offers advantages over RNAseq-only searches for cases where latency-associated viral transcripts are significantly less abundant than cellular transcripts. At present, DPS is more time-consuming than NGS and requires tissue-matched negative control samples. Thus, it should be seen an extension rather than a replacement for RNAseq analysis in virus discovery. DPS, however, has a critical advantage in making use of archival tumor FFPE tissues in which RNA is degraded. Development of a platform-independent human MS/MS reference database may markedly expand the potential for uncovering new human pathogens using DPS.

METHODS

Cell line, tissues, tissue microarray generation and immunohistochemistry

HEK293 cells (ATCC, Manassas, VA, USA) were maintained DMEM (10-013, Cellgro, Manassas, VA, USA) supplemented with 10% FBS (Sigma-Aldrich, St. Louis, MO, USA).

MCV positive, MCV negative MCC tumors were obtained from cooperative human tissue network (CHTN). Based on the MCV LT expression levels determined by CM2B4 staining, 11 MCV negative and 12 MCV positive tumors were selected for the differential Mass Spectrometry (dMS) study. Among those cores from 5 MCV positive, 4 MCV negative tumor FFPE blocks and a series of normal tissues (spleen, colon, brain, prostate, skin, adrenal gland, kidney, lung, uterus and tonsil) were used to generate a tissue microarray at UPMC Hillman Cancer Center Tissue and Research Pathology services.

Slides were deparaffinized in xylene and rehydrated in a series of ethanol solutions. Endogenous peroxidase activity was blocked by incubation of the slides with 3% hydrogen peroxide for 15 min. Epitope retrieval was performed using 1 mM EDTA buffer pH 8.0 at 125°C for 3 min, and 90°C for 15 sec in an antigen retrieval chamber (Decloaking chamber, Biocare medical, Pacheco, CA, USA). After blocking (Protein block, serum free, Dako, Ely, UK), monoclonal antibody CM2B4 generated by standard methods of immunizing mice with KLH-derivatized SRSRKPSSNASRGA peptide from the MCV T antigen (22) (0.6 µg/ml mAb, 1:1500,) and commercial antibodies CDKN2AIP (1:400, sc-81841, Santa Cruz, CA, USA), MASPIN/SERPINB5 (1:400 sc-271694), ATDC/TRIM29 (1:400 sc-376125) were diluted in (1 % BSA, 0.1 % gelatin, 0.5 % Triton-X, 0.05 % sodium azide in PBS pH 7.4) were applied to each section overnight at 4°C in a humidified chamber. Following extensive rinsing steps in TBS, sections were incubated with mouse Envision Polymer (Dako) for 30 min at room temperature, reacted with deaminobenzidine (DAB, Dako) and counterstained with hematoxylin (Dako). Images were acquired using Olympus microscope AX70 (Olympus Co., Ltd., Tokyo, Japan). All other chemicals were purchased from Sigma-Aldrich.

Sample selection, preparation for dMS

A total of 23 formalin-fixed, paraffin-embedded (FFPE) Merkel Cell carcinoma (MCC) tissue samples were selected on the basis of immunohistochemical staining that determined the presence/absence of MCV. The samples were anonymized to blind analysts from the MCV status until the proteomic sample preparation and mass spectrometric analysis were complete. Samples were sectioned to a 10 µm thickness using a microtome and stored on standard microscope slides.

Preparation of FFPE tissue for mass spectrometry

Deparaffinization was achieved with two xylene washes (3 min each), rehydrated with serial ethanol washes (100, 100, 95, and 70% for 1 minute each), and LC-MS grade water twice for 3 minutes each. After deparaffinization, 100 µl lysis buffer (300 mM Tris pH 8.0, 100 mM DTT, 2% SDS) was added to each tissue sample, followed by 30 min sonication, 1 h incubation at 95°C and 2 h incubation at 65°C. After centrifugation at 17,000 g for 10 min, the supernatants containing the extracted proteins were transferred to new eppendorf tubes and Pierce 660 nm Protein Assay kit with the IDCR packet (ThermoFisher Scientific, Waltham, MA, USA) was used to determine the total protein content.

Sample aliquots containing 30 µg of total protein were digested with trypsin using the Filter Aided Sample Preparation (FASP) protocol (30). In brief, the protein samples were added to YM30 Microcon microcentrifuge filters (Millipore, Darmstadt, Germany) and washed three times with 200 µl of urea buffer (100 mM Tris-HCl pH 8.0, 8 M urea), each with 15 min centrifugation at 14,000 g. Alkylation was performed by incubating at room temperatures for 20 minutes in 100 µl of urea buffer with 20 mM iodoacetamide. Samples were then washed 3 times with 100 µl urea buffer and then 3 times with 100 µl 50 mM ammonium bicarbonate, each with 10 min centrifugation at 14,000 g. 1.2 µg Sequencing Grade TPCK-treated trypsin (Promega, Fitchburg, WI, USA) were then added to each sample for overnight digestion in a humidified 37°C incubator. The resultant peptides were desalted using C18 Supelco cartridges (Supelco, Bellefonte, PA, USA), SpeedVac dried and then

reconstituted in 30 µl 0.1% formic acid for analysis. All other chemicals were purchased from Sigma-Aldrich.

Quality control samples were used to evaluate variability introduced by proteomic sample processing and mass spectrometric analysis. A set of 9 sample processing controls were created by combining equal amounts of undigested protein from the 23 extracted FFPE samples and processed alongside the experimental samples. A pool FFPE protein extract was divided into nine aliquots and processed together with the experimental samples to access sample preparation performance. A pooled instrument control sample was generated by combining equal volumes of all the digested samples and analyzed multiple times to monitor the stability of the mass spectrometer system over time. All sample identities were blinded to eliminate analyst bias and processed using a balanced block design to reduce variability introduced during sample processing and nLC-MS/MS analysis. The mean coefficient of variation (CV) for all quantified human peptides was used to characterize the biological (CV~90%) and technical (CV~30%) variation in the individual and replicate samples, respectively (**Supplemental Figure 1**).

Mass spectrometry and data processing

Complex mixtures of proteolytic peptides (0.2 µg for each injection) were analyzed by nano-flow liquid chromatography tandem mass spectrometry (nLC-MS/MS) with a nano Acquity UHPLC (Waters Corporation, Milford, MA) interfaced to a hybrid Orbitrap Velos Pro mass spectrometer (ThermoFisher Scientific). Peptide separation was carried out on a C18 PicoChip™ 25 cm column (New Objective, Woburn, MA) with a 66 min linear gradient of 2-35 % solvent B (acetonitrile/0.1% formic acid) at a 300 nL/min flowrate. The mass spectrometer was operated in positive ionization mode with an electrospray voltage of 1.9 kV and capillary temperature of 275°C. Ion sampling and accumulation was controlled with automatic gain control (AGC) and maximum injection time settings of 1,000,000 and 500 ms for full-scan high-resolution (MS1) mass spectra, and 5,000 and 100 ms for the low-resolution ion trap tandem (MS2) mass spectra, respectively. Data-dependent acquisition

recorded a full-scan MS1 spectrum at a resolution setting of 60,000 followed by 13 MS2 spectra at normalized collision energy setting of 35 with dynamic exclusion enabled. Separate nLC-MS/MS analyses that collect MS2 spectra on pre-defined precursor ions were performed using an isolation width of 2 m/z units and a relative collision energy setting of 35.

The raw mass spectrometry data were analyzed with MaxQuant software version 1.6.0.1 (24) that incorporates the Andromeda (25) protein identification search engine and label-free quantification tools. MS2 spectra were searched against the UniProt human proteome database (February 2013 release) using standard ORBItrap parameters and a reversed decoy database strategy that limits false peptide identifications rates to 1% or less. Briefly, a precursor mass tolerance setting of 20 and 4.5 ppm were used for the first and main database search, respectively. A mass tolerance setting of 0.5 Da was used for the MS2 fragment ions. Search enzyme specificity was defined as trypsin with maximum of two missed cleavages, fixed Cysteine carbamidomethylation, and variable methionine oxidation and protein N-terminal acetylation modifications. A minimum peptide length setting of 7 was used and the maximum number of modification per peptide was limited to 5. The “match between runs” and “matched unidentified” settings were enabled to prompt quantification of high-resolution MS1 features, regardless of the peptide sequence identification status.

Raw MS data files, together with MaxQuant quantification results, have been deposited to the ProteomXchange consortium via the MassIVE partner repository with the data set identifier PXD021520 and URL http://proteomecentral.proteomexchange. org/cgi/GetDataset?ID= PXD021520.

RNA extraction and SMART-library generation

Total RNA was isolated from MCV positive MCC tumor (R1165) using TRIzol (Ambion Inc, Austin, TX, USA) and was treated with TURBO DNase (ThermoFisher Scientific). RNA quality was examined by 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA) before (RIN value 5.3) and after ribosomal RNA depletion using RiboMinus Eukaryote kit

(ThermoFisher) according to the manufacturer's recommendations. Ribosome depleted samples were subsequently used for MCC-SMART library preparation. Libraries were prepared using the SMARTer PCR cDNA synthesis kit (Clontech, Mountain View, CA, USA) according to the manufacturer's recommendations with the following modifications: Switching Mechanism at 5' End of RNA Template (SMART) fusion primers were designed, which have the SMART sequence (5'-AAGCAGTGGTATCAACGCAGAGTAC-3') added to the 5' end of each dMS-identified MCV- or human-specific degenerate reverse primer listed in Supplementary Table 3. dMS-SMART-degenerate primer mix or a modified oligo(dT) primer (3' SMART CDS primer IIA) were used to prime first strand cDNA synthesis. Reaction mixtures consisting of 3.5 µl of RNA (~300 ng), 1 µl of 24 µM SMART primer mix (1.2 µM final concentration for each), or 1 µl of 12 µM 3'SMART CDS primer IIA (5'-AAGCAGTGGTATCAACGCAGAGTACT$_{(30)}$N$_{-1}$N-3' ,where N=A,C,G, or T and N$_{-1}$ =A,G, or C) were heated at 72°C for 3 min, and then the temperature was lowered to 47°C (0.1°C/min slope) for 2 min before the addition of 5.5 µl of master mix (2 µl of 5x first-strand buffer, 0.25 µl of 100 mM DTT, 1 µl of 10 mM dNTP mixture, 1 µl of 12 µM SMARTer IIA oligonucleotide, 0.25 µl of 40 U/µl RNase inhibitor, and 1 µl of 100 U/µl SMARTScribe RT). cDNA synthesis reaction mixtures of clinical specimens were incubated at 47°C for a total of 90 min, terminated at 70°C for 10 min, and brought to 4°C before the addition 0.1 µl RNAse H (5 U/µl, New England Biolabs, Ipswich, MA, USA). Reaction mixtures were incubated at 37°C for 20 min, subsequently kept at 4°C, and adjusted to 50 µl with water.

SMART cDNA was amplified by long-distance PCR on a thermocycler as follows using Advantage II reagents (Clontech): 7.5 µl SMART cDNA, 7.5 µl 10x Advantage 2 PCR buffer, 1.5 µl 50xdNTP mix (10 mM), 1.5 µl 5' PCR primer IIA (12 µM), 1.5 µl 50xAdvantage 2 polymerase mix, and 55.5 µl water (total of 75 µl). Reaction mixtures were cycled as follows: step 1, 95°C for 1 min; 35 cycles of step 2, 95°C for 15 s, 65°C for 30 s, and 68°C for 3 min followed by hold at 4°C.

Amplified SMART cDNA was purified with AMP-Pure magnetic beads (Beckman Coulter Genomics, Brea, CA, USA) using a ratio of 1.8x beads to sample according to the manufacturer's recommendations. Libraries were eluted in 30 µl of 10 mM Tris-Cl (pH 7.5) and then quantified on Agilent 2100 Bioanalyzer (Agilent) reagents.

NGS library generation, sequencing and analysis.

Nextera DNA Flex kit was used to generate libraries from SMART-cDNA templates following the manufacturer's instructions and sequencing was carried out on a NextSeq500 platform (Illumina Inc., San Diego, CA, USA) for 2x75 paired-end reads. Fastq files were imported into CLC Genomics Workbench 20.0 software (Qiagen, Hilden, Germany), paired-end reads 1 and 2 were merged, and duplicate reads were removed. Reads were filtered for Q-scores above 30, trimmed for quality (limit = 0.05) and ambiguity (2-nt maximum), and the illumina and SMART adaptor sequence were removed. Reads below 20 nt were discarded and paired-end reads were aligned to combined reference index from GRCh38 (Hg38) and MCV (JF813003) or to individual reference genomes. The following alignment settings were applied: mismatch = 2, insertion = 3, deletion = 3, length fraction = 0.8, and similarity fraction = 0.8. Sequencing data is deposited at NCBI GEO Platform accession number GSE157610.

Statistics

Feature selection was based on a combination of statistical significance and fold change difference. A two-sided equal variance Student's t-test on the log2 transformed intensities was used to determine the significance of difference between MCV positive and negative samples for all high-resolution MS1 features that consist of an "isotope group" without a corresponding human peptide sequence identification. Zero peptide intensity values were imputed with one-tenth of the global minimum of non-zero values to enable log2 transformation and fold change calculation. Unidentified MS1 features with at least 10 fold increase in MCV positive samples were ranked in an ascending order according to the Student's t-test p-values. 20 unidentified MS1 features with the highest significance were

subject to targeted nLC-MS/MS analysis. The targeted MS2 spectra were interpreted by manual de novo sequence analysis (31) (**Supplemental Figure 3-6**) and the identified sequence was confirmed with synthesized peptide standards. A representative select ion chromatogram depicts the relative abundance of Feature 1 peptide AYEYGPNPH(158)NSR in individual MCV positive and negative patient samples (**Supplemental Figure 7**).

Study approval

Tissues were obtained from cooperative human tissue network (CHTN) and examined under the University of Pittsburgh Institutional Review Board IRB 86-22: UPCI Tissue Banking Protocol.

AUTHOR CONTRIBUTIONS

T.T., Y.C., P.S.M., and N.A.Y. designed the experiments. T.T., P.S.C., Y.L., and X.Z performed the experiments. T.T. performed RNAseq, IHC, PCR and data analysis. P.S.C., M.S., and X.Z. performed differential mass spectrometry and related data analysis. N.A.Y., Y.C. and P.S.M supervised the project. T.T., P.S.C., X.Z., N.A.Y., Y.C., and P.S.M. wrote the paper. Y.C. and P.S.M. contributed equally to this work.

ACKNOWLEDGEMENTS

This project was supported by the National Institutes of Health (grant numbers R35 CA197463 to P. S. M. and CA170354 to Y. C.). P. S. M. and Y. C. are additionally supported as American Cancer Society Research Professors and receive support from the Pittsburgh Foundation (to P. S. M.) and the University of Pittsburgh Medical Center Foundation (to Y. C.). T.T. was supported in part by University of Pittsburgh Skin SPORE Career Enhancement Program Funding (NIH P50 CA121973-09) and Hillman Cancer Center Pilot Project Grant for Cancer Proteomics. Proteomics analysis was performed by the Hillman Cancer Center Proteomics Facility supported in part by award P30CA047904. This project used the University of Pittsburgh Health Sciences Sequencing Core at UPMC Children's Hospital of Pittsburgh for library generation and Illumina sequencing.

476 **REFERENCES**

477 1. Chang Y, Moore PS, and Weiss RA. Human oncogenic viruses: nature and
478 discovery. *Philos Trans R Soc Lond B Biol Sci.* 2017;372(1732).
479 2. Epstein MA, Achong BG, and Barr YM. VIRUS PARTICLES IN CULTURED
480 LYMPHOBLASTS FROM BURKITT'S LYMPHOMA. *Lancet (London,*
481 *England).* 1964;1(7335):702-3.
482 3. Blumberg BS, Alter HJ, and Visnich S. A "NEW" ANTIGEN IN LEUKEMIA
483 SERA. *Jama.* 1965;191:541-6.
484 4. Poiesz BJ, Ruscetti FW, Gazdar AF, Bunn PA, Minna JD, and Gallo RC.
485 Detection and isolation of type C retrovirus particles from fresh and cultured
486 lymphocytes of a patient with cutaneous T-cell lymphoma. *Proc Natl Acad*
487 *Sci U S A.* 1980;77(12):7415-9.
488 5. Dürst M, Gissmann L, Ikenberg H, and zur Hausen H. A papillomavirus
489 DNA from a cervical carcinoma and its prevalence in cancer biopsy
490 samples from different geographic regions. *Proc Natl Acad Sci U S A.*
491 1983;80(12):3812-5.
492 6. Choo QL, Kuo G, Weiner AJ, Overby LR, Bradley DW, and Houghton M.
493 Isolation of a cDNA clone derived from a blood-borne non-A, non-B viral
494 hepatitis genome. *Science.* 1989;244(4902):359-62.
495 7. Chang Y, Cesarman E, Pessin MS, Lee F, Culpepper J, Knowles DM, et al.
496 Identification of herpesvirus-like DNA sequences in AIDS-associated
497 Kaposi's sarcoma. *Science.* 1994;266(5192):1865-9.
498 8. Lisitsyn N, Lisitsyn N, and Wigler M. Cloning the differences between two
499 complex genomes. *Science.* 1993;259(5097):946-51.

9. Feng H, Shuda M, Chang Y, and Moore PS. Clonal integration of a polyomavirus in human Merkel cell carcinoma. *Science.* 2008;319(5866):1096-100.
10. Feng H, Taylor JL, Benos PV, Newton R, Waddell K, Lucas SB, et al. Human transcriptome subtraction by using short sequence tags to search for tumor viruses in conjunctival carcinoma. *J Virol.* 2007;81(20):11332-40.
11. Moore PS, and Chang Y. Why do viruses cause cancer? Highlights of the first century of human tumour virology. *Nature Reviews Cancer.* 2010;10(12):878-89.
12. Sarid R, Flore O, Bohenzky RA, Chang Y, and Moore PS. Transcription mapping of the Kaposi's sarcoma-associated herpesvirus (human herpesvirus 8) genome in a body cavity-based lymphoma cell line (BC-1). *J Virol.* 1998;72(2):1005-12.
13. Kwun HJ, da Silva SR, Shah IM, Blake N, Moore PS, and Chang Y. Kaposi's sarcoma-associated herpesvirus latency-associated nuclear antigen 1 mimics Epstein-Barr virus EBNA1 immune evasion through central repeat domain effects on protein processing. *J Virol.* 2007;81(15):8225-35.
14. Kwun HJ, Toptan T, Ramos da Silva S, Atkins JF, Moore PS, and Chang Y. Human DNA tumor viruses generate alternative reading frame proteins through repeat sequence recoding. *Proc Natl Acad Sci U S A.* 2014;111(41):E4342-9.
15. Toptan T, Fonseca L, Kwun HJ, Chang Y, and Moore PS. Complex alternative cytoplasmic protein isoforms of the Kaposi's sarcoma-associated

herpesvirus latency-associated nuclear antigen 1 generated through noncanonical translation initiation. *J Virol.* 2013;87(5):2744-55.

16. Tellam J, Sherritt M, Thomson S, Tellam R, Moss DJ, Burrows SR, et al. Targeting of EBNA1 for rapid intracellular degradation overrides the inhibitory effects of the Gly-Ala repeat domain and restores CD8+ T cell recognition. *J Biol Chem.* 2001;276(36):33353-60.

17. Rigatti LH, Toptan T, Newsome JT, Moore PS, and Chang Y. Identification and Characterization of Novel Rat Polyomavirus 2 in a Colony of X-SCID Rats by P-PIT assay. *mSphere.* 2016;1(6).

18. Toptan T, Yousem SA, Ho J, Matsushima Y, Stabile LP, Fernandez-Figueras MT, et al. Survey for human polyomaviruses in cancer. *JCI Insight.* 2016;1(2).

19. Meng F, Wiener MC, Sachs JR, Burns C, Verma P, Paweletz CP, et al. Quantitative analysis of complex peptide mixtures using FTMS and differential mass spectrometry. *J Am Soc Mass Spectrom.* 2007;18(2):226-33.

20. Wiener MC, Sachs JR, Deyanova EG, and Yates NA. Differential mass spectrometry: a label-free LC-MS method for finding significant differences in complex peptide and protein mixtures. *Anal Chem.* 2004;76(20):6085-96.

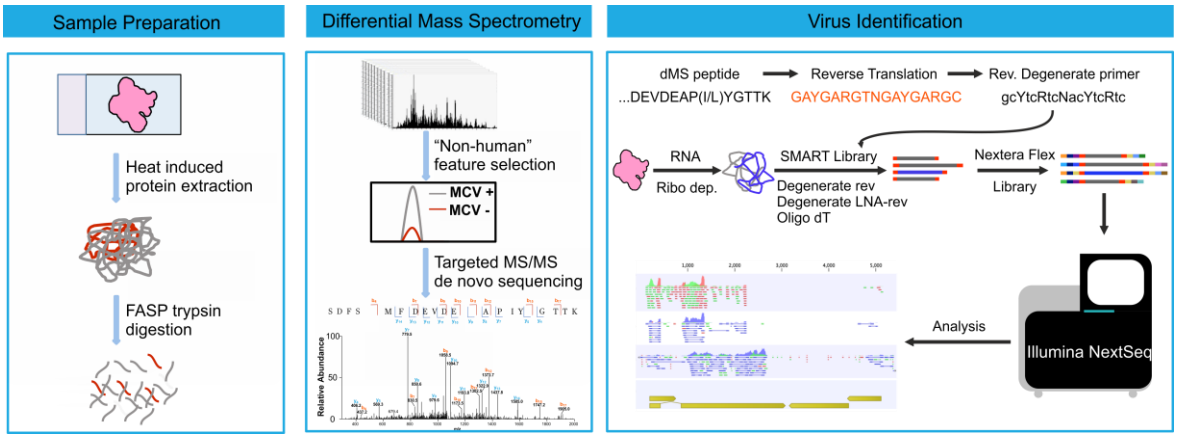
21. Zhao X, Deyanova EG, Lubbers LS, Zafian P, Li JJ, Liaw A, et al. Differential mass spectrometry of rat plasma reveals proteins that are responsive to 17beta-estradiol and a selective estrogen receptor modulator PPT. *J Proteome Res.* 2008;7(10):4373-83.

22. Shuda M, Arora R, Kwun HJ, Feng H, Sarid R, Fernandez-Figueras MT, et al. Human Merkel cell polyomavirus infection I. MCV T antigen expression

549 in Merkel cell carcinoma, lymphoid tissues and lymphoid tumors. *Int J*
550 *Cancer.* 2009;125(6):1243-9.
551 23. Harms PW, Harms KL, Moore PS, DeCaprio JA, Nghiem P, Wong MKK, et
552 al. The biology and treatment of Merkel cell carcinoma: current
553 understanding and research priorities. *Nat Rev Clin Oncol.*
554 2018;15(12):763-76.
555 24. Cox J, and Mann M. MaxQuant enables high peptide identification rates,
556 individualized p.p.b.-range mass accuracies and proteome-wide protein
557 quantification. *Nat Biotechnol.* 2008;26(12):1367-72.
558 25. Tyanova S, Temu T, and Cox J. The MaxQuant computational platform for
559 mass spectrometry-based shotgun proteomics. *Nat Protoc.*
560 2016;11(12):2301-19.
561 26. Berg MG, Yamaguchi J, Alessandri-Gradt E, Tell RW, Plantier JC, and
562 Brennan CA. A Pan-HIV Strategy for Complete Genome Sequencing. *J Clin*
563 *Microbiol.* 2016;54(4):868-82.
564 27. Berardi R, Morgese F, Onofri A, Mazzanti P, Pistelli M, Ballatore Z, et al.
565 Role of maspin in cancer. *Clin Transl Med.* 2013;2(1):8.
566 28. Masuda Y, Takahashi H, Sato S, Tomomori-Sato C, Saraf A, Washburn
567 MP, et al. TRIM29 regulates the assembly of DNA repair proteins into
568 damaged chromatin. *Nat Commun.* 2015;6:7299.
569 29. Yanagi T, Watanabe M, Hata H, Kitamura S, Imafuku K, Yanagi H, et al.
570 Loss of TRIM29 Alters Keratin Distribution to Promote Cell Invasion in
571 Squamous Cell Carcinoma. *Cancer Res.* 2018;78(24):6795-806.

572 30. Wiśniewski JR. Proteomic sample preparation from formalin fixed and
573 paraffin embedded tissue. *Journal of visualized experiments : JoVE.*
574 2013(79).
575 31. Kinter M, and Sherman N. *Protein Sequencing and Identification Using*
576 *Tandem Mass Spectrometry*. 2000:64-116.
577
578

FIGURES and FIGURE LEGENDS

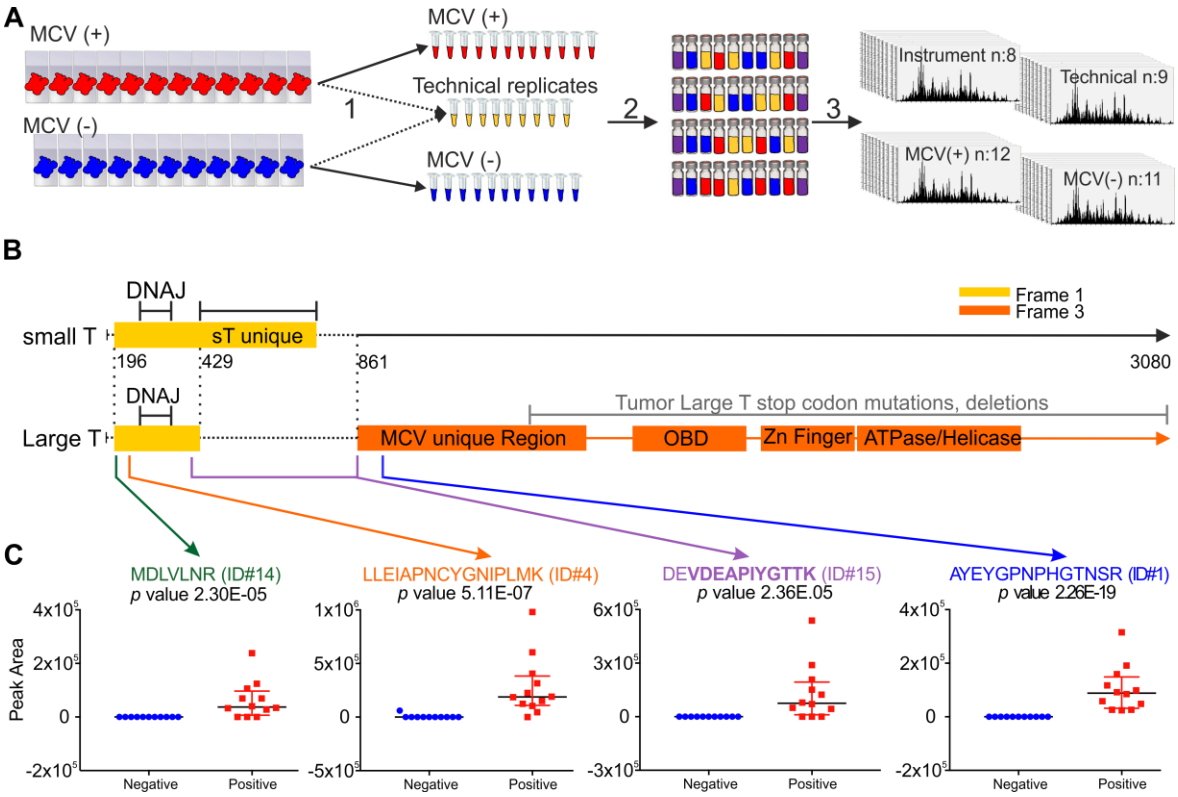


Graphical Abstract

Schematic representation of sample preparation workflow, mass spectrometry analysis and viral DNA sequence identification from proteomic sequence tags.

Figure 1

DPS can detect *de novo* the presence of a tumor virus

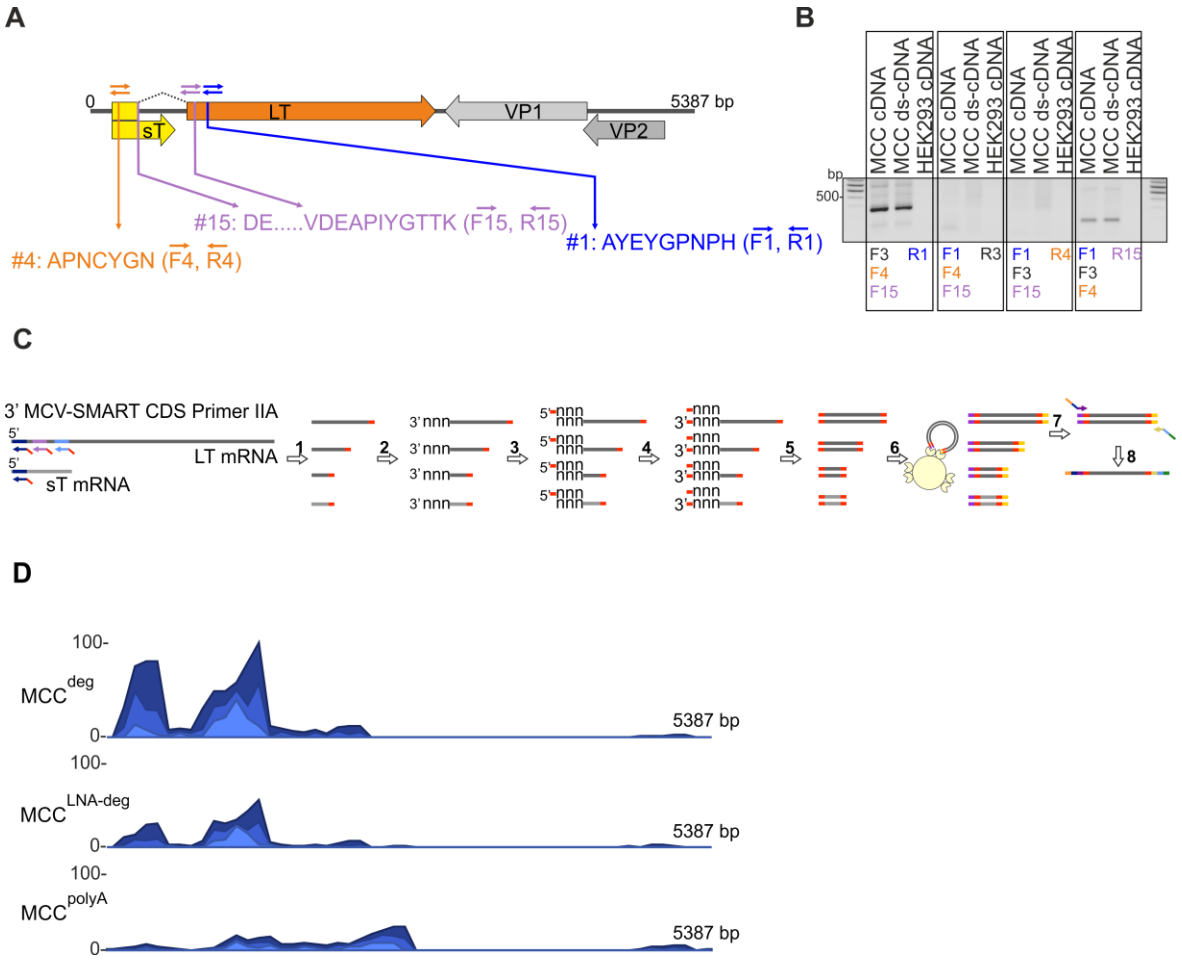


A. Workflow for dMS sample processing and instrumental analysis. 1. Deparaffinization, antigen retrieval and lysis: 10 µL from each sample (n:23) was combined and aliquoted into nine technical replicates. 2. FASP digestion step: Each sample was normalized to 30 µg. A total of 750 fmol of ovalbumin was added as an internal standard. A pooled instrument control was made by combining 5 µL from sample (n:32). Samples (n:33) were re-ordered. 3. nLC-MS/MS Analysis Step: Injection of ~0.2 µg on to C18 Picochip column Orbitrap Velos Pro and analysis. **B.** Schematic illustration of MCV T antigen transcripts. Small T (yellow-Frame 1) and Large T (yellow-Frame 1, orange-Frame 3) transcripts from the early region including start, splice and termination sites are shown. Both small t and Large T encode DNAJ domain. Small T and MCV unique domains, origin binding (OBD), zinc finger, ATPase and helicase domains are depicted. The location of mutations and deletions found in MCC tumor Large T are highlighted with a gray line. Positions of the four MCV peptides identified by dMS analysis are indicated with green, orange, purple, and blue arrows. **C.** Dot plots for the relative abundance of identified viral peptides in MCV positive (red, n:12) *vs.* negative (blue, n:11) MCC samples**.** Peptides and their rankings (Table 1) are shown in green (ID#14), orange (ID#4), purple (ID#15) and blue (ID#1).

606 Middle long lines indicates the mean and error bars represent standard deviation. P-
607 values were based on two-sided equal variance Student's t test.
608

Figure 2

dMS-identified peptides facilitate identification of viral sequences by NGS with cDNA libraries generated using degenerate oligos

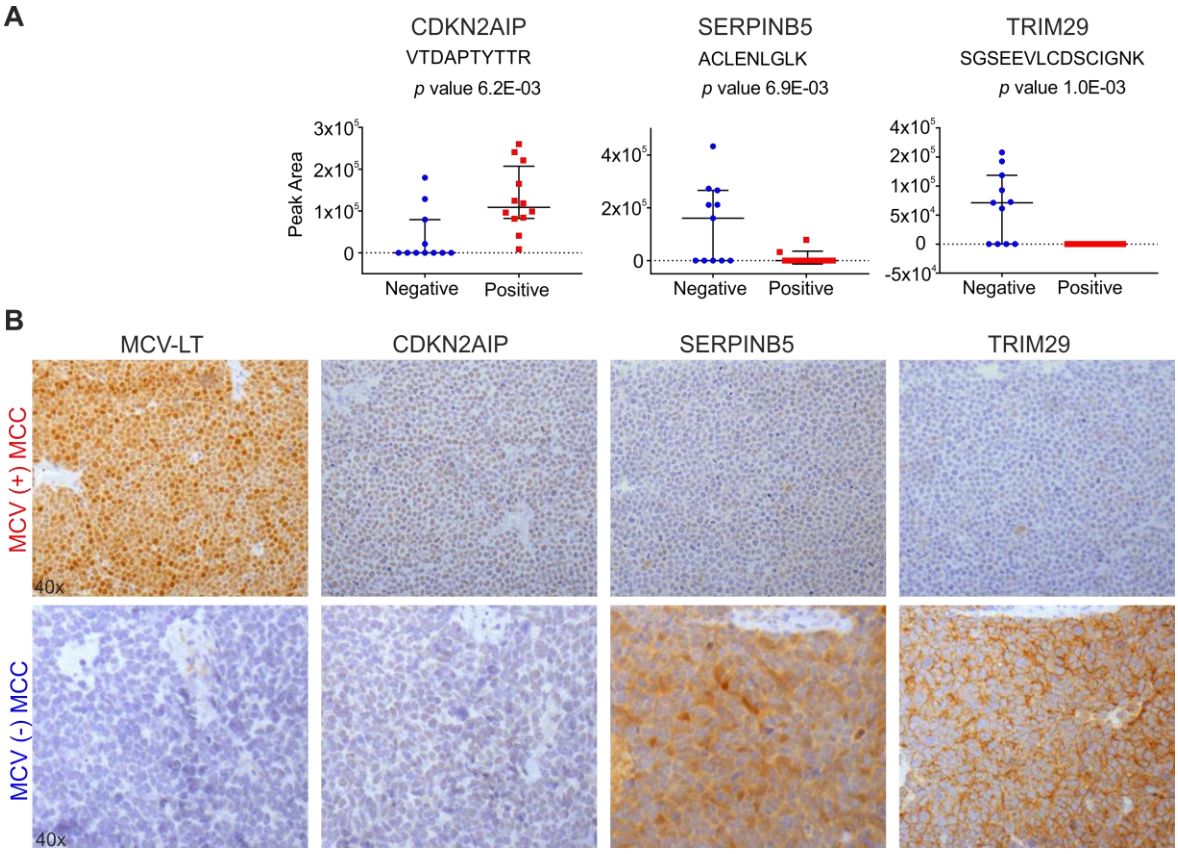


A. Schematic illustration of the MCV genome. Early (Large T (LT) yellow-orange, small T (sT) (yellow)) and late (VP1 (light gray), VP2 (dark gray)) region open reading frames are shown. The corresponding positions of the three MCV peptides identified by dMS (Features #4, #15 and #1) and degenerate primer binding sites are shown in orange, purple and blue arrows respectively. **B.** RNA extracted from MCC tissue (R16-67) or HEK293 cells were subjected to cDNA synthesis with random hexamers and additionally second strand synthesis for the MCC sample (ds-cDNA). cDNAs were amplified using the indicated combinations of degenerate primers (**Supplemental Table 3**) corresponding to the peptide sites highlighted in light blue (F1, R1), violet (F15, R15) and orange (F4, R4). F: forward, R: reverse. F3 and R3 (black) are non-MCV primers. **C.** Library generation using SMART oligos and Nextera-flex: **1**. 3' SMART CDS Primer IIA (Supplementary Table 4) mediated 1st strand synthesis. **2**. Tailing by RT. In the cDNA reaction, non-templated bases (nnn) are added to the ends of nascent cDNA by the terminal transferase

activity of reverse transcriptase (RT). **3.** SMARTer IIA oligo anneals to non-templated bases at cDNA ends (nnn). **4.** Template switch and extension at 3' end. The RT polymerase switches strands to transcribe the complement of the oligonucleotide, leaving the SMART adaptor at both ends of cDNA. **5.** Long-distance PCR with single 5' PCR Primer IIA amplifies libraries. **6.** Bead-linked transposomes mediate the simultaneous fragmentation of ds-cDNA and the addition of Illumina sequencing primers using Nextera-flex. **7.** Reduced-cycle PCR amplification amplifies sequencing ready DNA fragments and adds indexes and adapters. **8.** Sequencing-ready fragments are washed and pooled. **D. NGS coverage maps of MCC-RNAseq libraries.** RNAseq reads were obtained from three different samples to compare the efficiency of MCV read recovery using various primer pool sets for cDNA and library generation (**Supplementary Table 3**). Ribo-depteted MCC RNA (R11-65) was subjected to cDNA synthesis with SMART-degenerate oligo pool (MCC^{deg}), LNA modified SMART-degenerate oligo pool-SMART ($MCC^{LNA\text{-}deg}$), and modified oligo dT-SMART (MCC^{polyA}) and then subjected to library generation using Nextera-flex application. Standardized coverage depths (reads) for comparison purposes are indicated on the Y-axis for each alignment.

Figure 3

DPS can identify differentially expressed human peptides as potential biomarkers



A. Dot plots for the relative abundance of identified human peptides in MCV positive (red, n: 12) *vs.* negative (blue, n:11) MCC tumor samples. Middle long lines indicates the mean and error bars represent standard deviation. P values were based on two-sided equal variance Student's t test. **B.** Immunohistochemistry staining of MCC TMA. R10-115 and R15-03 are representative MCV positive (upper panel) and negative (lower panel) MCC cases respectively. According to the IHC staining results we detected SERPINB5 and TRIM29 in MCV negative cases and in none of the MCV positive cases as predicted by dMS analysis. MCV LT expression was detected using CM2B4 is a monoclonal antibody.

TABLES

Table 1. List of the top 20 significant proteomic features.

#ID	m/z	Charge	Mass	Calibrated retention time (min)	p value	Fold change	Peptide	Gene Name	Organism
1	521.571	3	1561.691	34.2	2.26E-19	4316.9	AYEYGPNP(TG, GT, AS, SA)SR	T antigen	MCV
2	390.542	3	1168.603	33.6	5.84E-09	3552.4	Unable to obtain amino acid sequence		N/A
3	565.309	1	564.302	35.2	1.99E-07	15.7	LQPVKcTGAR	PTTG1IP	Human/ Chimpanzee
4	923.484	2	1844.953	64.1	5.11E-07	50.6	XXEXA(PN, NP)cYGNXPXMK	T antigen	MCV
5	521.829	2	1041.643	57.4	6.65E-07	2770.5	DLIVATIAVK	ATIC	Human
6	579.612	3	1735.814	47.4	6.70E-07	2873.2	Unable to obtain amino acid sequence		N/A
7	454.926	3	1361.756	55.1	4.78E-06	10.4	Unable to obtain amino acid sequence		N/A
8	801.390	3	2401.147	64.7	2.02E-05	3048.5	(TX, XT)QFVDWY(SW, WS)EK		N/A
9	458.572	3	1372.694	31.0	2.06E-05	2028.4	Unable to obtain amino acid sequence		N/A
10	696.990	3	2087.950	48.6	2.09E-05	7199.7	NPSTVEAFDLAQSNSEHSR	PFAS	Human
11	406.229	3	1215.665	50.2	2.12E-05	4496.2	mKFNKK	U65	Human
12	449.743	2	897.471	43.2	2.23E-05	21636.5	AVLYNYR	C3	Human
13	738.883	2	1475.751	62.8	2.28E-05	41.5	DIINEEEVQFLK	AARS1	Human
14	451.742	2	901.469	63.5	2.30E-05	2583.7	(173)DXVXNR	T antigen	MCV
15	1076.475	2	2150.935	65.0	2.36E-05	5405.8	(714)DEVDEAPXYGTTK	T antigen	MCV
16	750.355	2	1498.695	51.9	2.46E-05	5616.0	STTSTIESFAAQEK	LUC7L3	Human
17	674.691	3	2021.051	68.9	2.48E-05	1711.9	VLFPGNSTQYNILEGLEK	MAP1B	Human
18	517.227	2	1032.440	25.2	2.70E-05	24.1	Unable to obtain amino acid sequence		N/A
19	565.326	3	1692.956	65.6	5.11E-05	13.2	QSXAEXXDXK		N/A
20	519.513	4	2074.024	31.7	2.11E-04	4433.1	Unable to obtain amino acid sequence		N/A

#ID, feature ID/Rank; m/z, monoisotopic mass to charge from MaxQuant (v1.6.0.1) output; Charge, the charge-state of the precursor ion; Mass, the predicted monoisotopic mass of the identified peptide sequence; Calibrated retention time (min), the recalibrated retention time in minutes in the elution profile of the precursor ion from MaxQuant (v1.6.0.1) output; p value, Student's t test p value after log2 transformation of peak area; Peptide, amino acid sequence associated with selected feature. X=isoleucine or leucine; c= Cysteine carbamidomethylated (+57.02); m= Methionine oxidation (+15.99).

Table 2. Summary of normalized read counts obtained from NGS analysis. Counts per million (CPM) reads following trimmed mean of M values (TMM) adjustment are calculated for each sample and gene. Transcripts per million (TPM) normalized values are indicated.

Sample	Gene	CPM-TMM adjusted	TPM
MCC-deg	sT	2,37	214,38
	LT	13,51	237,62
	VP1	0	0
	VP2	0,16	11,04
MCC-LNA.deg	sT	0,14	15,51
	LT	5,38	113,08
	VP1	0	0
	VP2	0,14	11,98
MCC-polyA	sT	0	0
	LT	2,1	95,96
	VP1	0	0
	VP2	0,11	19,31