Virology 513 (2018) 208–216

Contents lists available at ScienceDirect

Virology

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

Viral sequences in human cancer

Paul G. Cantalupo, Joshua P. Katz, James M. Pipas*

Department of Biological Sciences, University of Pittsburgh, Pittsburgh, PA 15260, USA

ARTICLE INFO

Keywords:
TCGA
Metagenomics
Cancer
Papillomavirus
Herpesvirus
Virome

ABSTRACT

We have developed a virus detection and discovery computational pipeline, Pickaxe, and applied it to NGS databases provided by The Cancer Genome Atlas (TCGA). We analyzed a collection of whole genome (WGS), exome (WXS), and RNA (RNA-Seq) sequencing libraries from 3052 participants across 22 different cancers. NGS data from nearly all tumor and normal tissues examined contained contaminating viral sequences. Intensive computational and manual efforts are required to remove these artifacts. We found that several different types of cancers harbored Herpesviruses including EBV, CMV, HHV1, HHV2, HHV6 and HHV7. In addition to the reported associations of Hepatitis B and C virus (HBV & HCV) with liver cancer, and Human papillomaviruses (HPV) with cervical cancer and a subset of head and neck cancers, we found additional cases of HPV integrated in a small number of bladder cancers. Gene expression and mutational profiles suggest that HPV drives tumorigenesis in these cases.

1. Introduction

Like all organisms, humans are constantly bombarded with microorganisms including viruses. Many diseases are the consequence of acute infection with viruses and in these cases the pathogen may be present for a limited time and be localized to specific tissues. Some viruses establish subclinical lifelong persistent or latent infections in their host thereby becoming part of the normal microbiome. Bacteriophages also form a major component of the human microbiome, their presence being indicative of their bacterial hosts. All species, including humans, must constantly respond to the myriad of endogenous viruses they harbor as well as to the transient presence of pathogenic viruses. Yet human viral ecology is poorly understood.

The Cancer Genome Atlas (TCGA) is a large database of deep sequencing of thousands of human tumors. This database has enabled the survey of viruses found in the tissue of cancer patients (Amirian et al., 2014; Cancer-Genome-Atlas-Research-Network, 2015, 2014a, 2014b; Kazemian et al., 2015; Khoury et al., 2013; Parfenov et al., 2014; Salyakina and Tsinoremas, 2013; Strong et al., 2013a, 2013b; Tang et al., 2013). Collectively these studies detected Human papillomavirus (HPV) sequences in nearly all cervical carcinomas as well as in a subset of squamous cell carcinomas of the head and neck; Hepatitis B and Hepatitis C viral sequences associated with a subset of liver cancers; and EBV gene expression in a subset of stomach cancers. Furthermore, these analyses detected viral associations with cancer that were previously unrecognized. For example, HPV was detected in a small number of bladder cancers and members of the Herpesvirus family were detected in some tumor and normal tissues. These studies provide an overview of the types of viruses present in human cancer and demonstrate the ability to identify molecular hallmarks associated with viral presence.

Oncogenic viruses contribute to tumorigenesis by expressing transforming proteins or ncRNAs that act on key cellular targets to alter cellular biology. In many cases the action of viral oncogenes results in the activation and repression of signaling pathways that are reflected in changes in cellular gene expression. In addition, integration of viral DNA is a hallmark of tumorigenesis for some viruses. Thus, specific changes in cellular gene expression patterns and/or viral integration events can be indicative of viral action driving tumorigenesis. In this manuscript, we report a survey of viral sequences present in TCGA data representing 22 distinct types of human cancers. This is the first study to combine DNA (WGS and WXS) and RNA sequencing data sets to search for viral sequences present in human cancers and to deduce their effects of cellular gene expression.

2. Results

2.1. Virus detection pipeline and removal of artifacts

To identify known viruses present in tumor or normal tissue from cancer patients we compared sequences in TCGA databases to the reference genomes for all known viral species in NCBI (Viral RefSeq). Unmapped reads from whole genome sequencing (WGS), whole exon sequencing (WXS), and RNA-seq libraries were obtained from TCGA BAM files. High quality reads were selected and aligned with Bowtie 2

* Corresponding author.
E-mail address: pipas@pitt.edu (J.M. Pipas).

http://dx.doi.org/10.1016/j.virol.2017.10.017
Received 10 August 2017; Received in revised form 10 October 2017; Accepted 19 October 2017

Table 1
Number of TCGA patients, samples and libraries processed.

		Patients				Samples							
Cancer Abbr	**Cancer**	**Virus positive**	**Total Patients**	**RNA-Seq**	**DNA-Seq**	**Total Samples**	**Tumors**	**Normals**	**RNA-Seq[a]**	**WXS**	**WGS**	**Both RNA & DNA[b]**	**Total Libraries[c]**
BLCA	Bladder Urothelial Carcinoma	20 (7.5%)	268	267	253	541	271	270	289	526	145	274	960
BRCA	Breast invasive carcinoma	0 (0.0%)	49	49	48	100	55	45	100	91	4	91	195
CESC	Cervical squamous cell carcinoma	252 (98.8%)	255	252	208	467	256	211	256	420		209	676
COAD	Colon adenocarcinoma	74 (18.2%)	407	407	398	429	408	21	429 (431)	402	69	417	902
GBM	Glioblastoma multiforme	2 (1.2%)	162	162	162	169	169		169 (339)	167	4	167	510
HNSC	Head and Neck squamous cell carcinoma	125 (24.2%)	517	500	516	1096	518	578	542 (544)	1078	15	524	1637
KICH	Kidney Chromophobe	2 (3.0%)	66	66	66	91	66	25	91	91		91	182
KIRC	Kidney renal clear cell carcinoma	0 (0.0%)	50	50	50	100	50	50	100	97		97	197
KIRP	Kidney renal papillary cell carcinoma	0 (0.0%)	78	78	77	101	76	25	101 (107)	100		100	207
LAML	Acute Myeloid Leukemia	0 (0.0%)	173	173	71	173	173		173	71		71	244
LGG	Brain Lower Grade Glioma	1 (1.0%)	100	100	94	100	100		100	94		94	194
LIHC	Liver hepatocellular carcinoma	25 (35.2%)	71	71	67	104	68	36	104	96		96	200
LUAD	Lung adenocarcinoma	0 (0.0%)	57	57	57	114	59	55	114 (115)	108		108	223
LUSC	Lung squamous cell carcinoma	0 (0.0%)	56	56	56	99	56	43	99 (100)	93		93	193
OV	Ovarian serous cystadenocarcinoma	0 (0.0%)	93	93	86	100	100		100	86		86	186
PAAD	Pancreatic adenocarcinoma	2 (5.0%)	40	40	40	41	40	1	41	41		41	82
PRAD	Prostate adenocarcinoma	0 (0.0%)	56	56	56	100	56	44	100	93		93	193
READ	Rectum adenocarcinoma	32 (20.5%)	156	156	154	162	157	5	162	159	32	160	353
SKCM	Skin Cutaneous Melanoma	2 (2.0%)	98	98	98	100	100		100	100		100	200
STAD	Stomach adenocarcinoma	36 (25.9%)	139	139	137	143	127	16	143	141		141	284
THCA	Thyroid carcinoma	0 (0.0%)	66	66	62	132	74	58	132	116		116	248
UCEC	Uterine Corpus Endometrial Carcinoma	0 (0.0%)	95	95	93	100	95	5	100	98		98	198
Totals			**3052**			**4562**	**3074**	**1488**	**3545 (3727)**	**4268**	**269**	**3267**	**8264**

[a] Sample number equals number of analysis IDs processed (BAM files) except where number of analysis IDs is given in parenthesis.
[b] Number of samples that have both RNA and DNA data.
[c] Number of analysis IDs where each corresponds to a unique BAM file.