Detection and quantification of viral RNA in human tumors using open source pipeline: viGEN

Krithika Bhuvaneshwar[1], Lei Song[1], Subha Madhavan[1], Yuriy Gusev[1]

kb472@georgetown.edu, polosong@gmail.com, sm696@georgetown.edu, yg63@georgetown.edu

[1]Innovation Center for Biomedical Informatics, Georgetown University, Washington DC 20007, USA

Present Address of all authors: 2115 Wisconsin Ave NW, Suite 110, Washington DC 20007 USA.

Corresponding author: Yuriy Gusev, Krithika Bhuvaneshwar

ABSTRACT

Introduction

An estimated 17% of cancers worldwide are associated with infectious causes. The extent and biological significance of viral presence/infection in actual tumor samples is generally unknown but could be measured using human transcriptome (RNA-seq) data from tumor samples.

We present an open source bioinformatics pipeline viGEN that combines existing well-known and novel RNA-seq tools for not only detection and quantification of viral RNA, but also variants in the viral transcripts.

Methods

The pipeline includes 4 major modules: The first module allows to align and filter out human RNA sequences; second module maps and count (remaining un-aligned) reads against reference genomes of all known and sequenced human viruses; the third module quantifies read counts at the individual viral genes level thus allowing for downstream differential expression analysis of viral genes between experimental and controls groups. The fourth module calls variants in these viruses. To the best of our knowledge, there are no publicly available pipelines or packages that would provide this type of complete analysis in one open source package.

Results

In this paper, we use this pipeline in a case study to examine viruses present in RNA-seq data from 75 TCGA liver cancer patients. We were able to quantify viral transcriptomes at a viral-gene/CDS level, find differentially expressed viral transcripts between the groups of patients, extract variants, and connect them to clinical outcome. The results presented corresponded with published literature in terms of rate of detection, viral gene expression patterns and impact of several known variants of HBV genome. Results also show novel information about distinct patterns of expression and co-expression in Hepatitis B, Hepatitis C, Human Endogenous Retrovirus (HERV) K113 viruses.

Conclusion

This pipeline is generalizable, and can be used to provide novel biological insights into the significance of viral and other microbial infections in complex diseases, tumorigeneses and cancer immunology. The source code, with example data and tutorial is available at: https://github.com/ICBI/viGEN/.

Keywords – RNA-seq, viral detection, liver cancer, TCGA, variant analysis, next-generation sequencing, cancer immunology

BACKGROUND

Hepatocellular carcinoma (HCC, the primary malignancy of the liver) is now the third most common cancer in the world affecting more than half a million people. The incidence of liver cancer varies greatly by race and ethnicity; and about 3 times more common in men than women [1]. The most common type of HCC - caused by Hepatitis B and C viruses (HBV and HCV) are most prevalent in Asia and Africa, as the presence of virus predisposes people to liver disease and subsequently HCC [2]. In such high prevalence areas of the world, HBV infection is often acquired at birth or in early childhood. In the US, Asian American, Native Hawaiian and Pacific Islanders (AANHPI) account for more than 50% of people infected with HBV, although many of

them are unaware that they harbor the virus [1]. Infection can also occur in high risk groups like injection drug users and health care workers [1, 3].

Apart from HBV and HCV infections, other risk factors for liver cancer include heavy alcohol consumption, obesity, diabetes, tobacco smoking, certain rare genetic conditions, or cirrhosis (scarring of the liver) [1]. Liver disease triggered by obesity and diabetes is called nonalcoholic fatty liver disease (NAFLD) [2]. Even though these risk factors are known, it's not clearly understood how these normal liver cells become cancerous [4].

Existing methods of screening and treatments

For patients who are at high risk of liver cancer, screening using ultrasound, and also a blood test for alpha-fetoprotein (AFP, protein made by the liver and yolk sac of the developing fetus and normally found in fetal blood) is done every 6-12 months [5]. Currently, detection of Hepatitis B is done using serology tests and involves measurement of HBV specific antigen and antibody markers that identify different phases of infection or immunity. Presence of Hepatitis (Hep) B surface antigen (HBsAg) or Hepatitis B type e antigen (HBeAg) indicates chronic and infectious hepatitis B infection; antibody HBsAg (Anti-HBs) presence indicates immunity through vaccine or a past Hep B infection; presence of antibody to hepatitis B core antigen (Anti-HBc) indicates an ongoing or past Hep B infection; presence of IgM antibody to Hep B core antigen (IgM anti-HBc) indicates recent acute infection with hepatitis B [6, 7].

Detection of Hepatitis C virus is typically done using Enzyme immunoassay (EIA) to detect hepatitis C antibody; or Hepatitis C RNA assays are used to determine the viral load. Genetic testing is then done to check for the type of Hep C infection, which can be of six types (genotypes 1 through 6). In the US, baby boomers (born 1945-1965) are encouraged to get tested, as they may be treated with antiviral drugs to prevent progression to cancer [8].

Vaccine is available to protect against HBV infection, but not for HCV. Even though vaccines exist, it can only protect against infections if they are administered before the person is exposed to the cancer-promoting virus [9]. Recent advances in screening have helped in early detection of

cirrhosis. In recent times, quantitative assays for HBsAg and HBeAg are also being used in identifying patients likely to respond to anti-HBV treatments, although more work is needed to standardize these new assays [10]. When early detection is successful, treatments include surgery to remove part of the liver (partial hepatectomy) or liver transplant [1]. Other treatments include cryosurgery and radiation therapy for cases where cancer has not metastasized. For patients where cancer has metastasized, targeted therapy, chemotherapy or clinical trials are tried. So early detection of liver disease is crucial, but has been challenging since the symptoms may not appear until the cancer has advanced [4].

Viral mechanisms of action

HBV is an enveloped partially double-strand DNA virus in the hepadnavirus family, and is a known oncogenic virus. In a HBV infection, the virus enters the bloodstream and infects the liver cells by active viral replication. During this time, the HBV genome integrates into the host chromosome and becomes the basis for chronic infection. HBV DNA integration may offer a selective growth advantage on infected cells and promote tumor progression. The integration sites of HBV DNA usually occur in genes involving growth control or cell signaling. When HBV DNA integrates into the host, chromosomal instability is also increased (e.g., large deletions, amplification, and translocations). Integrated HBV DNA sequences are found in about 80% of human Hep B induced liver cancer [11].

HCV is an enveloped single-strand RNA virus in the flavivirus family. Unlike HBV, HCV does not have a reverse transcriptase, so is not able to integrate into the genome of hepatocytes. HCV causes liver cancer by an indirect pathway when it produces chronic inflammation, cell death, proliferation and cirrhosis in the liver. Acute HCV infection patients that have large CD4+ and CD8+ T cell response in their blood are known to have a better chance of recovery. In contrast, patients who lack T cell response seems to indicate that the patient will develop chronic HCV disease [11].

Opportunities with Next generation sequencing

The popularity of next-generation sequencing (NGS) technology has exploded in the last decade. NGS technologies are able to perform rapid sequencing, and in a massively parallel fashion [12]. In recent years, applications of NGS technologies in clinical diagnostics have been on the rise [13, 14]. Amongst the various NGS technologies, whole-transcriptome sequencing, also called RNA-seq has been very popular with methods and tools being actively developed. Exploring the genome using RNA-seq gives a different insight than looking at the DNA since the RNA-seq would have captured actively transcribed regions. Every aspect of data output from this technology is now being used for research, including detection of viruses and bacteria [15-17]. They are also independent of prior sequence information, and require less starting material compared to conventional cloning based methods, making it a powerful and exciting new technology in virology [12].

Our pipeline viGEN combines existing and novel RNA-seq tools to not only detect and quantify read counts at the individual viral genes level, but also detect viral variants from human RNA-seq data. The input file to our pipeline is a fastq [18] file, so our viGEN pipeline can be extended to work with genomic data from any NGS technology. Our pipeline can also be used to detect and explore other types of microbes as long as the sequence information is available in NCBI [19].

There are a number of existing pipelines that detect viruses from human transcriptome data. Of these, very few pipelines offer quantification at the gene/CDS level. A comprehensive comparison of these pipelines is provided in Table 1.

Table1: Comparison of existing pipelines that detect viruses from human transcriptome data

Tool Name	Detect viruses from Human RNA-seq data	Perform quantification at viral-gene/CDS level	Works on DNAseq, RNAseq or Both	Variant calling at viral-variant level	Discover viral integration sites	Other comments
Virana [20]	Yes	Identifies microbial transcripts, does not quantify	Both	No	Yes	Also offers analysis of homologs

VirusSeq [21]	Yes	No	Both	No	Yes	Pre-designed to work on a select set of 18 viruses
Viral Fusion Seq [22]	Yes	No	Both	No	Yes	Can also detect fusion events
Virus Finder [23]	Yes	No	Both	No	Yes	Can be applied to samples infected with undiagnosed viruses
PathSeq [24]	Yes	No	Both	No	No	
RINS [25]	Yes	No	RNA-seq	No	No	Generates contigs with these non-human sequences
viGEN (our pipeline)	Yes	Yes	Both	Yes	No	

Our goal was not to compete with these other tools, but to offer a convenient and complete end–to-end publicly available pipeline to the bioinformatics community. To the best of our knowledge there are no publicly available pipelines or packages that would provide this type of complete analysis in one package. Customized solutions have been reported in the literature however were not made public.

Our pipeline incorporates existing best practices and tools available, and we used novel tools only when there was no other option. The results presented in this paper are a proof of concept of our pipeline. In addition, we have made available an end-to-end tutorial demonstrated on a publicly available RNA-seq sample from NCBI SRA (http://www.ncbi.nlm.nih.gov/bioproject/PRJNA279878), providing step-by-step instructions on the analysis steps, results of analysis, along with the code at https://github.com/ICBI/viGEN/. Our plan is to package this pipeline and make it open source through Bioconductor [26], allowing users to perform analysis on either their local computer or the cloud.

METHODS

In this paper, we used RNA-seq data from a publicly available human liver cancer data from the TCGA collection [27]. The raw genomic data was downloaded after obtaining special access from NCBI dbGAP (http://www.ncbi.nlm.nih.gov/gap). Existing well-known and novel RNA-seq tools were used to detect and quantify viral RNA at the genome and gene/CDS level. Once the viral genomes were detected, it allowed for downstream differential expression analysis of viral genes

between experimental and controls groups. The viral variants detected in our pipeline can also give more insight into the mutations in these viruses and their impact on the tumor.

The data used for analysis in this paper consisted of 75 liver cancer patients in the TCGA data collection. The cohort includes three sub-types - 25 patients infected with Hepatitis B virus (labelled as 'HepB'), 25 patients infected with Hepatitis C virus (labelled as 'HepC') and 25 patients that are co-infected with Hepatitis B and C viruses (labelled as 'HepB+HepC'). These sub-type classifications were defined based on 'Viral Hepatitis Serology' attribute from the clinical information.

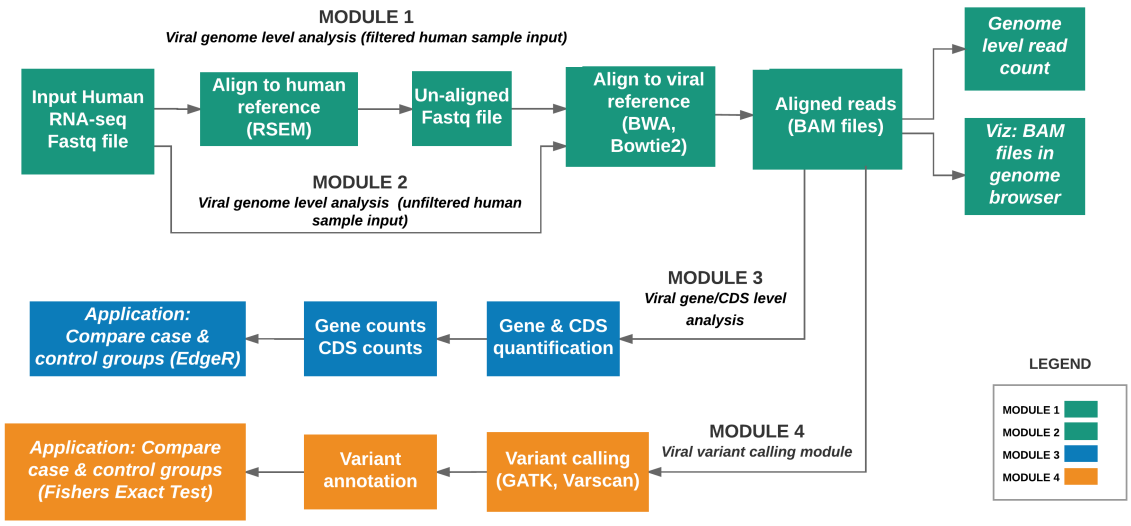
We were interested in exploring all viruses existing in humans. So we first obtained reference genomes of all known and sequenced human viruses obtained from NCBI [19] (as of Sep 2015), and merged them into one file (referred to as the 'viral reference file') in fasta file format [28].

The viGEN pipeline includes 4 major modules:

Module 1: Viral genome level analysis (filtered human sample input)

In Module 1 (labelled as 'filtered human sample input'), the human RNA sequences were aligned to the human-reference genome using RSEM [29] tool using the Globus Genomics platform [30]. The RSEM tool filters out all the sequences the aligned to the human genome, and outputs all sequences that did not align to the human genome (hence the name 'filtered human sample input'). These un-aligned sequences were taken and aligned to the 'viral reference file' using popular alignment tools BWA [31] and Bowtie2 [32]. Figure 1 shows an image of our viGEN pipeline.

Figure 1: viGEN pipeline. Each module has a color, shown in the legend



Module 2: Viral genome level analysis (unfiltered human sample input)

In Module 2 (labelled as 'unfiltered human sample input'), the human RNA seq sequences were directly aligned to the 'viral reference' using BWA and Bowtie2 using the Globus genomics platform without any filtering.

The reason for using two methods to obtain the viral genomes in human RNA-seq data (Module 1 and Module 2) was to allow us to be as comprehensive as possible in viral detection.

The aligned reads from Module 1 and 2 were in the form of BAM files [33], from which read counts were obtained for each viral genome species (referred to as 'genome level counts') using Samtools idxstats [34] and Picard BAMIndexStats [35] tools. Only those virus species that had average genome count more than a minimum threshold (set to 100 reads) across samples in each sub-group (Hep B, Hep C, HepB+HepC) were selected for the next step of the pipeline.

Once the viral genomes were detected, it allowed us to examine them through a genome browser. We also checked the genome level counts to see if the Hepatitis B and C viruses were detected from this output, and compared it with the information from viral serology tests.

Module 3: Viral gene/CDS level analysis

The BAM files from Module 1 and 2 (from Bowtie2 and BWA) were input into the Module 3 (referred to as 'viral gene/CDS level analysis'), which calculated quantitate read counts at the individual viral genes level. We found existing RNA-seq quantification tools to be not sensitive enough for viruses, and hence developed our own algorithm for this module. Our in-house algorithm used region-based information from the general-feature-format (GFF) files [36] of each viral genome, and the reads from the BAM file. It created a summary file, which had a total count of reads within or on the boundary of each region in the GFF file. This is repeated for each sample and for each viral GFF file. At the end, a matrix is obtained where the features (rows) are regions from the GFF file, and the columns are samples.

The read count output from Module 3 (viral gene/CDS module) allowed for downstream differential expression analysis of viral genes between experimental and controls groups. In this pilot study, we examined the differences between "Dead" and "Alive" samples at the viral-transcript level for each sub-group using Bioconductor tool EdgeR [37] in R (http://www.R-project.org). Cox proportional hazards (Cox PH) regression model [38] was also applied to look at overall survival time and event in the Hepatitis B sub-group.

Module 4: Viral variant calling module

The BAM files from Module 1 and 2 (from Bowtie2) were also input to Module 4 to detect mutations in these viruses (referred to as 'viral-variant calling module'). The BAM files were first sorted coordinate-wise using Samtools [34]; PCR duplicates were removed using tool Picard [35], then the chromosomes in the BAM file were ordered in the same way as the reference file using Picard. The Viral reference file was created from combining all known and sequenced human viruses obtained from NCBI [19] (as of Sep 2015). Popular variant calling tool GATK's HaplotypeCaller and UnifiedGenotypeCaller [39] functions were used to detect variants. Another variant calling tool, Varscan2 [40] known for detecting low-frequency variants [41], was also used. The current version of our pipeline uses Varscan2. Low quality and low depth variants were flagged, but not filtered out, in case these low values were attributed to low viral load. Once the

variants were obtained, they were merged to form a multi-sample VCF file. Only variants that had a variant in at-least one sample were retained. PLINK [42] was used to perform case-control association test (Fishers Exact Test) to compare 'alive' and 'dead' samples in the HepB and HepC groups.

Tutorial in Github

Since access to TCGA raw data is controlled access, we could not use this dataset to create a publicly available tutorial. So we looked for publicly available RNA-seq dataset to demonstrate our pipeline with an end-to-end workflow. We chose one sample (SRR1946637) from publicly available liver cancer RNA-seq dataset from NCBI SRA (http://www.ncbi.nlm.nih.gov/bioproject/PRJNA279878). This dataset is also available through EBI SRA (http://www.ebi.ac.uk/ena/data/view/SRR1946637). The dataset consists of 50 Liver cancer patients in China, and 5 adjacent normal liver tissues. We downloaded the raw reads for one sample, and applied our viGEN pipeline to it. A step-by-step workflow that includes – description of tools, code, intermediate and final analysis results are provided in Github: https://github.com/ICBI/viGEN/. This tutorial has also been provided as Additional File 1.

RESULTS

Detection of Hepatitis B and C viruses at the genome level

We used our viGEN pipeline to get genome-level read counts obtained from viruses detected in the RNA of human liver tissue. We then checked to see if Hepatitis B and C viruses were detected from this output, and how it compared with the information from viral serology tests. We know from the serology test (obtained from the clinical data) that of the 75 samples in this pilot study, 25 have Hepatitis B; another 25 have Hepatitis C and the rest if the samples have both Hepatitis B and C.

We were able to detect RNA from HepB and HepC at the genome level in corresponding samples of HCC patients. Table 2A and 2B shows a comparison of the viral detection from

serology (blood) with viral detection from RNA-seq data (tumor tissue) for Hepatitis B and C respectively.

Table 2A (top) and 2B (bottom): Comparison of viral detection from serology (blood) with viral detection from RNA-seq data (tumor tissue) for Hepatitis B and Hepatitis C respectively. The results shown are from Module 1 ('Filtered human sample input') performed using Bowtie2 alignment tool

	Serology from blood		**TOTAL**
RNA-seq from tumor tissue	HepB Negative	HepB Positive	
HepB Negative	25	40	65
HepB Positive	0	10	10
Total	25	50	75

	Serology from blood		**TOTAL**
RNA-seq from tumor tissue	HepC Negative	HepC Positive	
HepC Negative	7	12	19
HepC Positive	18	38	56
Total	25	50	75

For Hepatitis B detection, we used a cut off read count of 1000 to define a 'HepB positive' state. Using this cut-off, the Hep B virus was correctly detected in only 10 of the 50 samples (i.e. only 10 samples had read counts more than 1000). The rest of the 40 patients in our cohort had read counts between 0-1000 and were grouped as 'HepB negative'. The 25 HepC patients were correctly identified as 'HepB negative' (Table 2A).

The RNA-serology comparison for Hepatitis C detection was more complicated - the serology test only tests for presence of any Hep C antibody; on the other hand, Hepatitis C genomes can be of 6 types – Genotype 1 to 6. So we looked at the read counts in any of these 6 Hepatitis C genotypes. We used the first quartile as the threshold for a positive detection (i.e. 'Hep C Positive'). Using this cut-off, the Hep C virus was correctly detected in 38 of 50 samples, and out of the 25 patients that did not have Hep C, 7 samples were correctly identified as having Hep C (Table 2B).

Only those virus species that had average genome count more than a minimum threshold (set to 100 reads) across samples in each sub-group (Hep B, Hep C, HepB+HepC) were selected for the next step of the pipeline (Modules 3 and 4). In addition to Hepatitis B and C, several other

viruses came up in this short list including Human endogenous retrovirus K113 (HERV K113), sub-types of Human herpes virus, Human papillomavirus, Human adenovirus and others. A complete list is provided in Additional File 2.

Landscape of viruses at the genome level

We explored the genome level BAM files (from Module 1) through the IGV genome browser to see if we could find any interesting patterns. We looked at the landscape of HBV genome, HCV genomes and HERV K113 genome across all the samples.

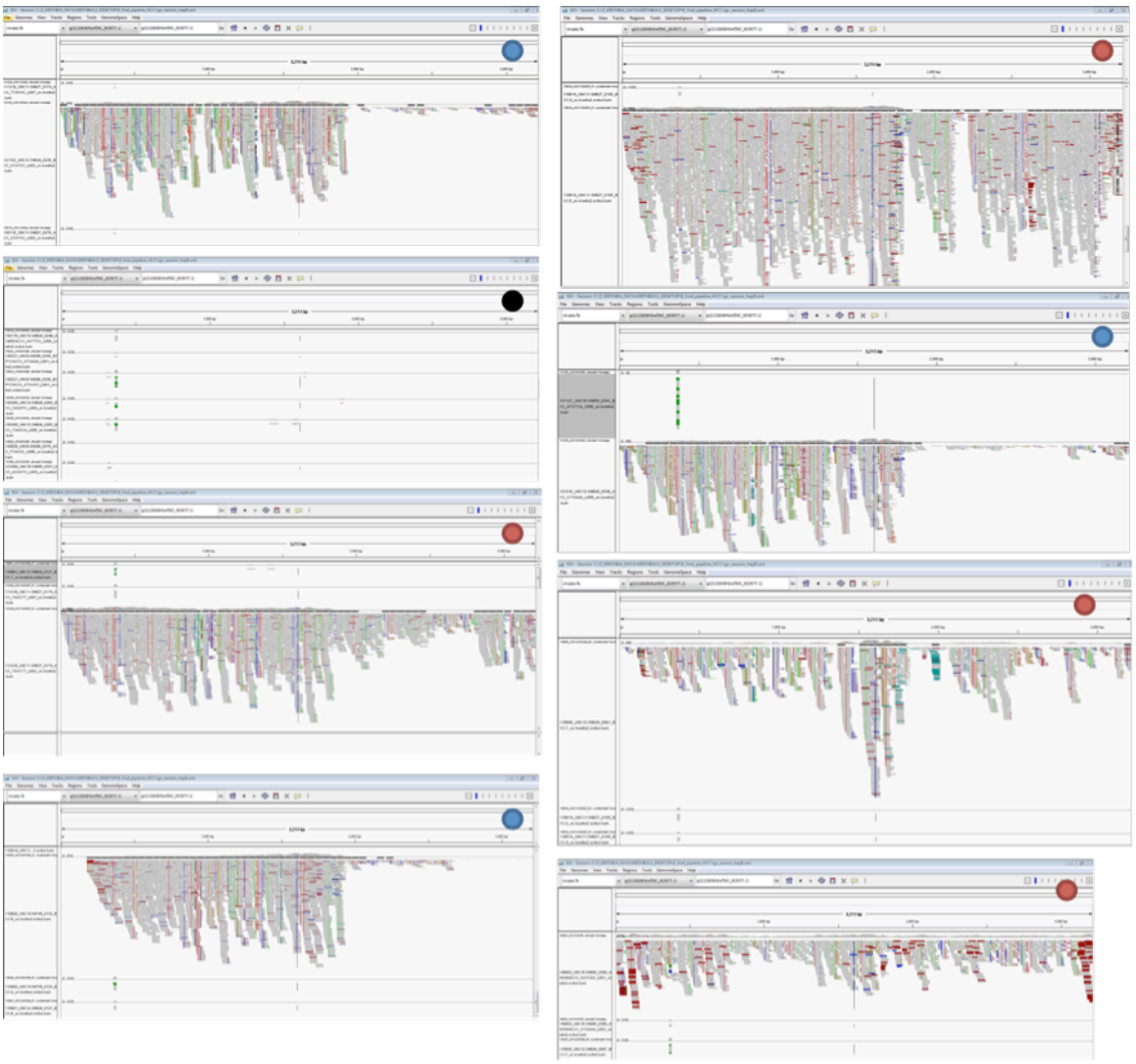
In the landscape of HBV, we found the samples could be grouped under one of three patterns (Figure 2):

(a) Pileups seen throughout the HepB genome (labelled as 'full')

(b) Pileups seen only on the left half of the HepB genome (labelled as 'truncated') and

(c) Very few or no pileups seen (labelled as 'empty').

Figure 2: Landscape of HBV genome across the 25 HepB samples. The BAM files used were from Module 1 generated using Bowtie2. The image shows the three types of patterns of pileup distribution namely full (red dot), truncated (blue dot), and empty (black dot).



We also explored these patterns w.r.t clinical variable Dead/Alive status (Table 3). In general, patients who are 'Alive' seem to have more pileups that those who 'Died' although this conclusion cannot be strongly made since there are only a small number of samples in each pattern.

Table 3: Summary of genome browser pattern for Hepatitis B genome with Dead/Alive status

	HepB samples (25 samples total)		
Pattern Name	**Meaning**	**Alive**	**Dead**
Empty	Very few or no pileups seen	12	6
Full	Pileups seen throughout the HepB genome	3	1
Truncated	Pileups seen only on the left half of the HepB genome	1	2
	Total	**16**	**9**

In the landscape of HCV (NC_004102.1), we found two types of patterns (a) terminal repeats seen on the right end of the genome (labelled as 'terminal repeats') and (b) terminal repeats seen on the right end of the genome along with pileups seen in other parts of the genome (labelled as 'Terminal repeats + some pileup'). For other HCV types (Hepatitis C virus genotype 2, 3 and 6) the pattern consisted of only the 'terminal repeats'. HCV is an RNA virus belonging to the Flaviviridae family. The high error rate of RNA-dependent RNA polymerase and other factors have driven the evolution of HCV into 7 different genotypes and more than 67 subtypes. The quasispecies variations of the HCV genome and high rate of recombination causes large variability and diversity [43]. Also its unusual pattern of terminal repeats might be partially due to the fact that those regions are highly evolutionary conserved [44, 45] and do not show much variability as opposed to coding regions. This shows the general difficulty and complexity in the detection of this virus using standard RNA-seq. Transcription in these viruses is very transient and hard to capture [46].

In the landscape of HERV K113 virus, we found three types of patterns (a) Pileups seen throughout the genome (labelled as 'full'), Terminal repeats seen at both ends of the genome, sparse pileup in other regions (labelled as 'medium') and Terminal repeats at both ends of the genome (labelled as 'terminal repeats').

The figures in Additional File 3 shows these above mentioned pattern types. We also matched these patterns to Dead/Alive status and summarized our findings in the tables in Additional File 3.

The range patterns in these viral landscapes indicate that information can be extracted in a meaningful way from the read information, and it adds to the validity of our approach.

Comparing 'dead' and 'alive' samples in the HepB subgroup using viral gene/CDS data

To get a more detailed overview of the viral landscape, we examined the human RNA-seq data to detect and quantify viral gene/CDS regions. We then examined the differences between 'Dead' and 'Alive' samples at the viral-transcript level on the Hepatitis B sub-group.

Out of 25 patients, 16 were alive (baseline group), and 9 dead (comparison group) as per the clinical data. The significant results are shown in Table 4.

Table 4: Differential expression analysis of transcript level read counts on Hep B sub-group comparing Dead and Alive samples. These results shown used the viral-gene/CDS data obtained from Module 1 (using alignment tool Bowtie2) + Module 3. The table shows results with q value < 0.06 and sorted based on LogFC in the descending order.

Name of region (Name of virus_region_start position of region_end position of region)	Log Fold change (logFC)	Log counts per million (logCPM)	P Value	Q Value (FDR)	Name of virus	Region annotation
NC_001405.1_intron_9724_12307	2.527	6.463	4.02E-08	1.22E-05	Human mastadenovirus C	Gene=L1, locus_tag=HAdVC_gp10, note=precedes capsid protein precursor pIIIa CDS
NC_001405.1_intron_9724_11039	2.5	6.451	6.65E-08	1.28E-05	Human mastadenovirus C	Gene=L1, locus_tag=HAdVC_gp10, note=precedes encapsidation protein 52K CDS
NC_001405.1_gene_10866_11023	2.5	6.451	6.65E-08	1.28E-05	Human mastadenovirus C	Gene=VAII, locus_tag=HAdVC_gs02, GeneID:2653002
NC_001405.1_transcript_10866_11023	2.5	6.451	6.65E-08	1.28E-05	Human mastadenovirus C	Gene=VAII, locus_tag=HAdVC_gs02, GeneID:2653002
NC_001405.1_exon_10866_11023	2.5	6.451	6.65E-08	1.28E-05	Human mastadenovirus C	Gene=VAII, locus_tag=HAdVC_gs02, GeneID:2653002
NC_001405.1_intron_10580_14015	2.428	6.475	9.24E-08	1.64E-05	Human mastadenovirus C	Gene=E2B, locus_tag=HAdVC_gp04
NC_003977.1_gene_1814_2452	1.128	13.449	1.71E-06	0.000243	Hepatitis B virus	Contains Gene C that produces pre-code protein external core antigen; HBeAg. HBeAg is produced by proteolytic processing of the pre-core protein
NC_003977.1_CDS_1814_2452	1.128	13.449	1.71E-06	0.000243	Hepatitis B virus	Contains Gene C that produces pre-code protein external core antigen; HBeAg.
NC_003977.1_CDS_1901_2452	0.828	12.42	0.000507	0.053928	Hepatitis B virus	Contains Gene C, encodes core antigen HBcAg
NC_022518.1_STS_7174_7323	-0.992	9.122	0.000309	0.034527	Human endogenous retrovirus K113	Sequence-tagged site (STS), locus_tag =Q779_gp1, standard_name=D6S2277, UniSTS:59918
NC_022518.1_STS_5100_5381	-1.051	9.532	0.000118	0.0139	Human endogenous retrovirus K113	Sequence-tagged site (STS), standard_name= D22S1651, UniSTS: 474031
NC_022518.1_region_1112_6746	-1.186	13.022	3.49E-06	0.000463	Human endogenous retrovirus K113	gag-pro-pol; two -1 frameshifts predicted to occur to produce a fusion protein; the location of frameshifts has not been determined

NC_018464.1_region_1_927	-1.288	12.784	5.78E-07	9.45E-05	Shamonda virus	mol_type=genomic RNA, isolate=Ib An 5550, taxon:159150, segment=S
NC_003977.1_CDS_155_835	-2.121	16.335	1.67E-15	1.11E-12	Hepatitis B virus	Encodes Gene S that produces small envelope protein, S protein; S glycoprotein; S-HBsAg,
NC_003977.1_gene_2307_4838	-2.133	12.655	2.61E-15	1.11E-12	Hepatitis B virus	Gene P, encodes protein P
NC_003977.1_CDS_2307_4838	-2.133	12.655	2.61E-15	1.11E-12	Hepatitis B virus	Gene P, encodes protein P
NC_003977.1_CDS_3205_4050	-2.352	8.67	1.93E-12	6.84E-10	Hepatitis B virus	Gene S, encodes middle envelope protein pre-S2/S
NC_002645.1_gene_293_20568	-2.598	6.126	3.93E-05	0.00491 1	Human coronavirus 229E	locus_tag=HCoV229Egp1, GeneID: 918764, replicase polyprotein 1ab
NC_003977.1_gene_2848_4050	-2.75	11.741	5.84E-22	6.20E-19	Hepatitis B virus	Encodes Gene S that produces a large surface protein/L glycoprotein/L-HBsAG
NC_003977.1_CDS_2848_4050	-2.75	11.741	5.84E-22	6.20E-19	Hepatitis B virus	Encodes Gene S that produces a large surface protein/L glycoprotein/L-HBsAG

From the differential expression analyses, the two most informative results were (1) a region of the Hepatitis B genome that produced the HBeAg and HBcAg proteins were overexpressed in the 'dead' patients and (2) another region of the Hepatitis B genome that produced HBsAg protein was overexpressed in the 'alive patients.

In detail, we saw several important findings as described below:

(a) Region NC_003977.1_CDS_1814_2452 of the Hepatitis B genome was 2.18 times overexpressed (log fold change = +1.128) in 'dead' patients. This region contains Gene C that produces pre-code protein external core antigen; HBeAg. HBeAg is produced by proteolytic processing of the pre-core protein

(b) Region NC_003977.1_CDS_1901_2452 which was 1.74 times overexpressed (log fold change = +0.8, FDR = 0.053) in 'dead' patients contains Gene C as above, but encodes a different core antigen HBcAg

(c) Region NC_003977.1_CDS_2848_4050 of the Hepatitis B genome was 6.73 times over expressed (log fold change = -2.7) in the 'alive' patients of compared to the 'dead' patients. This region encodes Gene S that produces a large surface protein/L glycoprotein/L-HBsAG

(d) We also found several regions of the Human endogenous retrovirus K113 (HERV K113) viral genome (NC_022518.1_region_1112_6746, NC_022518.1_STS_5100_5381 and

NC_022518.1_STS_7174_7323) to be about 2 times overexpressed on average in 'alive' patients (log fold change = -1.186, -1.051, -0.992).

Survival analysis (Cox Regression) using viral gene/CDS data

Based on the results from previous section, we selected two most informative regions from the Hepatitis B genome (log counts per million from NC_003977.1_CDS_2848_4050, NC_003977.1_CDS_1814_2452) for a Cox Proportional Hazard (Cox PH) model to look at overall survival event and time. This model was applied on the 25 Hep B and 25 HepB+HepC samples to maximize power. The result from this model (Table 5), are consistent with the results from differential expression analysis:

(a) The Cox PH model shows that assuming other covariant to be constant, unit increase in expression of this region NC_003977.1_CDS_1814_2452, increases the hazard of event (death) by 70%.

(b) On the other hand, that assuming other covariant to be constant, unit increase in expression of this region NC_003977.1_CDS_2848_4050, decreases the hazard of event (death) by 43%.

(c) The overall model is significant with p-value < 0.05 from the Log rank test (also called Score test).

Table 5: Cox proportional hazard survival analysis (across 25 HepB samples and 25 HepB + HepC Samples). These results shown used the viral-gene/CDS data obtained from Module 1 (using alignment tool Bowtie2) + Module 3. *Coef: coefficient (Beta) of the model; exp(coef): Hazard Ratio; se(coef) : Standard Error; Pr(>|z|) : P-value*

Formula:
coxph(formula = survObject ~ NC_003977.1_CDS_2848_4050 + NC_003977.1_CDS_1814_2452)

Results from the model:
n= 37, number of events= 5
(13 observations deleted due to missingness)

Covariate	**coef**	**exp(coef)**	**se(coef)**	**Z**	**Pr(>|z|)**
NC_003977.1_CDS_2848_4050	-0.5548	0.5742	0.7434	-0.746	0.456
NC_003977.1_CDS_1814_2452	0.5302	1.6993	0.6145	0.863	0.388

Covariate	**exp(coef)**	**exp(-coef)**	**Lower 0.95**	**Upper 0.95**
NC_003977.1_CDS_2848_4050	0.5742	1.7415	0.1337	2.465
NC_003977.1_CDS_1814_2452	1.6993	0.5885	0.5096	5.667

Concordance= 0.654 (se = 0.188)
Rsquare= 0.12 (max possible= 0.329)
Likelihood ratio test= 4.74 on 2 df, p=0.09343
Wald test = 0.75 on 2 df, p=0.6856
Score (logrank) test = 10.58 on 2 df, p=0.00503

Comparing 'dead' and 'alive' samples in the HepB subgroup using viral-variant data

We performed variant calling on the viral data to see if it can add valuable information to the tumor landscape in humans. We then compared the 'dead' and 'alive' samples at the viral-variant level on the 25 patients in the Hepatitis B sub-group. For this analysis, the outputs from both from Module 1 and 2 were fed into Module 4.

Most of the top variants from filtered human sample (Module 1 + Module 4) (Additional File 4: Table S3-A) and unfiltered human sample (Module 2 + Module 4) (Additional File 4: Table S3-B) using variant caller Varscan2, were the same. We collated the significant common results (p value <= 0.05) in Table 6. Among these results, we saw several missense and frameshift variants in Gene X of the Hepatitis B genome (nucleotide 1479), Gene P (2573, 2651, 2813), and a region that overlaps Gene P and PreS1 (nucleotides 2990, 2997, 3105, 3156). All these variants were found mutated more in the cases than controls. Other significant common results included variants in Gene C (nucleotide 1979, 2396) and variants in PreS2 region (nucleotide positions 115, 126 and 148).

In addition, there were two missense variants that were common among the top results, but not significant (p value = 0.06). They were variants in the X gene of the Hepatitis B genome (nucleotides 1762 and 1764).

Among the significant common results to both, were a few variants of the Human endogenous retrovirus K113 complete genome (HERV K113). These include nucleotide positions 7476, 7426 and 8086. These map to frameshift and missense mutations in the putative envelope protein of this virus (Q779_gp1, also called 'env').

Table 6: Results of case-control association test applied on the results from viral variant calling (showing only common results between two possible analysis steps). The table is sorted based on Annotation. Annotation includes gene name, protein name, etc., separated by commas, multiple annotations separated by semi-colon

CHR (Chromosome)	Species (Name of Virus)	BP (Base pair)	A1 (minor allele)	C_A (Number of cases with A1)	C_U (number of controls with A1)	A2 (major allele)	P (P value)	Annotation from GFF file

gi|21326584|ref|NC_003977.1|	Hepatitis B virus	1479	C	4	0	A	0.02857	Gene=X, product=X protein, protein_id=NP_647606.1
gi|21326584|ref|NC_003977.1|	Hepatitis B virus	2573	C	0	6	T	0.03571	Gene=P, product=polymerase, protein_id=NP_647604.2
gi|21326584|ref|NC_003977.1|	Hepatitis B virus	2651	T	4	0	C	0.00476	
gi|21326584|ref|NC_003977.1|	Hepatitis B virus	2813	C	2	0	T	0.03571	
gi|21326584|ref|NC_003977.1|	Hepatitis B virus	2990	T	2	0	A	0.02222	Gene=P, product=polymerase, protein_id=NP_647604.2; Gene=S, product=large envelope protein, protein_id=YP_355333.1
gi|21326584|ref|NC_003977.1|	Hepatitis B virus	2997	C	2	0	T	0.03571	
gi|21326584|ref|NC_003977.1|	Hepatitis B virus	3105	C	2	0	A	0.02222	
gi|21326584|ref|NC_003977.1|	Hepatitis B virus	3156	G	4	0	A	0.00476	
gi|21326584|ref|NC_003977.1|	Hepatitis B virus	1979	G	2	0	A	0.03571	Gene=C, product=pre-capsid protein, protein_id=YP_355335.1, NP_647607.1
gi|21326584|ref|NC_003977.1|	Hepatitis B virus	2396	0	4	0	CG	0.01499	
gi|21326584|ref|NC_003977.1|	Hepatitis B virus	115	C	2	0	A	0.02222	Pre S2 region, ID=id0, Dbxref=taxon: 10407, Is_circular=true, gbkey=Src, genome=genomic, mol_type=genomic DNA, strain=ayr
gi|21326584|ref|NC_003977.1|	Hepatitis B virus	126	C	2	0	T	0.02222	
gi|21326584|ref|NC_003977.1|	Hepatitis B virus	148	G	2	0	A	0.02222	
gi|21326584|ref|NC_003977.1|	Hepatitis B virus	1762	T	0	4	A	0.06061	Gene=X, Name=NP_647606.1, product=X protein, protein_id=NP_647606.1
gi|21326584|ref|NC_003977.1|	Hepatitis B virus	1764	A	0	4	G	0.06061	
gi|548558394|ref|NC_022518.1|	Human endogenous retrovirus K113	7476	0	10	14	TACTG	0.00600	ID=gene0, Name=Q779_gp1; ID=cds0, Name=YP_008603282.1, product=putative env, protein_id=YP_008603282.1
gi|548558394|ref|NC_022518.1|	Human endogenous retrovirus K113	7426	G	3	0	A	0.00714	
gi|548558394|ref|NC_022518.1|	Human endogenous retrovirus K113	8086	T	3	0	C	0.00714	

DISCUSSION

Detection of Hepatitis B and C viruses at the genome level

We used our viGEN pipeline to get genome-level read counts obtained from viruses detected in the RNA of human liver tissue. In our results, HBV was detected in 20% of the samples. This is in concordance with earlier analyses of TCGA liver cancer cohort study [16, 47], which detected the HBV virus in 23% and 32% (with typically low counts range) of cases respectively.

It has also been reported that the viral gene X (HBx) was the most predominately expressed viral gene in liver cancer samples [47] which is in concordance with our findings where the peak number of reads were observed for gene X region of the HBV genome (see Figure 2).

We also compared the HBV and HCV detection from our data with the viral serology tests (Table 2A and 2B). We see some, but not a lot of concordance. There could be several reasons for the differences we see between RNA from tissue and serology:

(a) We should remember that we are looking at viruses detected in RNA of human cancer tissue, and it is well known that this landscape is different from blood (which is used for the serology tests) or normal liver. According to [48], lab tests prove that HBV DNA replication and HBsAg are generally detected in different hepatocytes, while HBV DNA replication is generally, but not consistently seen in hepatocytes with HBcAg.

(b) If the viral DNA is integrated into the host (seen in acute infection stages and often precedes tumor development), in spite of having antigen/antibody markers in blood (causing serology test to be positive), it will not produce any RNA particles, causing low viral load in RNA [48, 49].

(c) The tumor site acts like a viral reservoir, which allows the virus to accumulate and be stable, and allows for replication of virus. This makes the virus hard to detect through serology, and might be detectable when examining the tumor site [50-52].

(d) A patient could be in a 'HBV carrier state', which is characterized by the presence of HBsAg in the serum, low or undetectable levels of HBV DNA, normal aminotransferase activity and lack of HBeAg [53]. That means that in this stage, low levels of HBV DNA could cause low viral load in RNA even through the serology test is positive. In this stage, use of immunosuppressive therapies can lead to reactivation of infection [53].

(e) Serology tests are known to be un-reliable when the immune system becomes dysfunctional and may also explain the false positives seen in the results [54, 55].

These results show that even though the genome-level viral counts detected through human RNA-seq are not a 100% match to viral serology data, it gives a good overview of the viral

landscape in the tumor sample, and demonstrates the complex dynamics of infection and immune response. These results also indicate that a deeper look inside these viruses is warranted.

Comparing 'dead' and 'alive' samples in the HepB subgroup using viral gene/CDS data

To get a more detailed overview of the viral landscape, we examined the human RNA-seq data to detect and quantify viral gene/CDS regions. We then examined the differences between 'Dead' and 'Alive' samples at the viral-transcript level on the Hepatitis B sub-group (Table 4).

From the differential expression analyses, the two most informative results were (1) a region of the Hepatitis B genome that produced the HBeAg protein was overexpressed in the 'dead' patients and (2) another region of the Hepatitis B genome that produced HBsAg protein was overexpressed in the 'alive patients.

Presence of HBeAg or HBcAg is an indicator of active viral replication; this means the person infected with Hepatitis B can likely transmit the virus on to another person. Typically, loss of HBeAg is an indicator of recovery from acute Hepatitis B infection. Active viral replication could allow the virus to persist in infected cells, and increase the risk of disease [56, 57]. So our results, showing that antigens HBeAg and HBcAg were overexpressed in 'dead' patients compared to 'alive patients' makes sense, indicating that these patients never recovered from acute infection.

The results also indicate a higher level of HBsAg in the 'alive' patients compared to the 'dead' patients. The highest levels of HBsAg in the virus are known to occur in the 'immunotolerant phase'. This pattern is seen in patients who are inactive carriers of the virus i.e. they have the wild type DNA, and the virus has been in the host for so long, that the host does not see the virus as a foreign protein in the body, and hence there's no immune reaction against the virus. In this phase, there is known to be minimal liver inflammation and low risk of disease progression [58-60]. This could explain why we saw higher level of HBsAg in the 'alive' patients compared to the 'dead' patients.

Also among the significant results were three regions from the Human endogenous retrovirus K113 (HERV K113) genome (with negative log fold change) that were overexpressed in the 'alive' patients. Two of these regions were Sequence-tagged sites (STS) and the third region was in the gag-pro-pol region that has frameshifts. HERV could protect the host from invasion from related viral agents through either retroviral receptor blockade or immune response to the undesirable agent [61].

Overall, we found that our results from viral-gene/CDS level make biological sense, with much of the results validated through published literature.

Comparing 'dead' and 'alive' samples in the HepB subgroup using viral-variant data

We performed variant calling on the viral data to see if it can add valuable information to the tumor landscape in humans. We then compared the 'dead' and 'alive' samples at the viral-variant level on the 25 patients in the Hepatitis B sub-group.

Among the significant results (Table 6) included variants in Gene C (nucleotide 1979, 2396) and variants in PreS2 region (nucleotide positions 115, 126 and 148). The Gene C region creates the pre-capsid protein, which plays a role in regulating genome replication [62]. The mutation in the 2396 position lies in a known CpG island (ranging from 2215-2490), whose methylation level is significantly correlated with hepatocarcinogenesis [63]. Mutations in PreS2 are associated with persistent HBV infection, and emerge in chronic infections. The PreS1 and PreS2 regions are known to play an essential role in the interaction with immune responses because they contain several epitopes for T or B cells [64].

Mutations in the 1762/1764 positions of the X gene are known to be associated with greater risk of HCC [64] [65], and is independent of serum HBV DNA level [65]. This mutation combination is also known to be associated with hepatitis B related acute-on-chronic liver failure [66]. It is predicted that mutations associated with HCC variants are likely generated during HBV-induced pathogenesis. The A1762T/G1764A combined mutations was shown to be a valuable

biomarker in the predicting the risk of HCC [64] [65]; and are often detected about 10 years before the diagnosis of HCC [64].

Among the significant common results to both, were a few variants of the Human endogenous retrovirus K113 complete genome (HERV K113). These variants map to frameshift and missense mutations in the putative envelope protein of this virus (Q779_gp1, also called 'env'). Studies have shown that this envelope protein mediates infections of cells [67]. HERV K113 is a provirus and is capable of producing intact viral particles [68]. Studies have shown a strong association between HERV-K antibodies and clinical manifestation of disease and therapeutic response [69] [70]. It is hypothesized that retroviral gene products can be 'reawakened' when genetic damage occurs through mutations, frameshifts and chromosome breaks. Even though the direct oncogenic effects of HERVs in cancer are yet to be completely understood, it has shown potential as diagnostic or prognostic biomarkers and for immunotherapeutic purposes including vaccines [70].

CONCLUSION

With the decreasing costs of NGS analysis, our results show that it is possible to detect viral sequences from whole-transcriptome (RNA-seq) data in humans. Our analysis shows that it is not easy to detect DNA and RNA viruses from tumor tissue, but certainly possible. We were able to not only quantify them at a viral-gene/CDS level, but also extract variants. Our goal is to facilitate better understanding and gain new insights in the biology of viral presence/infection in actual tumor samples. The results presented in this paper using the 75-sample dataset from TCGA are in correspondence with published literature and are a proof of concept of our pipeline.

This pipeline is generalizable, and can be used to examine viruses present in genomic data from other next generation sequencing (NGS) technologies. It can also be used to detect and explore other types of microbes in humans, as long as the sequence information is available from the National Center for Biotechnology Information (NCBI) resources.

This pipeline can thus be used on cancer and non-cancer human NGS data to provide additional insights into the biological significance of viral and other types of infection in complex diseases, tumorigeneses and cancer immunology. We are planning to package this pipeline and make it open source to the bioinformatics community through Bioconductor.

LIST OF ABBREVIATIONS

HBV- Hepatitis B virus,
HCV – Hepatitis C Virus,
HERV K113 – Human Endogenous Retrovirus K113,
TCGA – The Cancer Genome Atlas,
HCC - Hepatocellular carcinoma
NAFLD - nonalcoholic fatty liver disease
Hep B - Hepatitis B
Hep C - Hepatitis C
HepB + HepC - coinfected with both Hepatitis B and C virus
HBsAg - Hepatitis B surface antigen
HBeAg - Hepatitis B type e antigen
NGS - next-generation sequencing
RNA-seq - whole transcriptome sequencing
BAM - Binary version of Sequence alignment/map format
CDS – coding sequence
Cox PH - Cox Proportional Hazard
HBx - viral gene X
STS - Sequence-tagged sites
NCBI - National Center for Biotechnology Information
GFF - general-feature-format

DECLARATIONS

Availability of data and material

- The TCGA liver cancer dataset was used in the analysis and writing of this manuscript. The data can be obtained from https://cancergenome.nih.gov/

- Since access to TCGA raw data is controlled access, we could not use this dataset to create a publicly available tutorial. So we looked for publicly available RNA-seq dataset to demonstrate our pipeline with an end-to-end workflow. We chose one sample (SRR1946637) from publicly available liver cancer RNA-seq dataset from NCBI SRA (http://www.ncbi.nlm.nih.gov/bioproject/PRJNA279878). This dataset is also available through EBI SRA (http://www.ebi.ac.uk/ena/data/view/SRR1946637). The dataset consists of 50 Liver cancer patients in China, and 5 adjacent normal liver tissues. We downloaded the raw reads for one sample, and applied our viGEN pipeline to it. A step-by-step workflow that includes – description of tools, code, intermediate and final analysis results are provided in Github: https://github.com/ICBI/viGEN/.

Project details:

Project name: viGEN
Project home page: https://github.com/ICBI/viGEN/
Operating system(s): The R code is platform independent. The shell scripts can run on Unix, Linux, or iOS environment
Programming language: R, bash/shell

Other requirements: N/A
License: N/A
Any restrictions to use by non-academics: N/A

Competing interests
The authors do no have any competing interests

Funding
Not applicable

Author contributions
KB and YG designed the pipeline. KB and LS implemented the pipeline. KB and YG wrote the manuscript with editorial comments from SM.

Acknowledgements
Not applicable

REFERENCES

1. **Cancer Facts and Figures 2016** [http://www.cancer.org/acs/groups/content/@research/documents/document/acspc-047079.pdf] Accessed: April 22, 2016
2. **Hepatocellular Carcinoma** [http://emedicine.medscape.com/article/197319-overview] Accessed: April 22, 2016
3. **Cancer Facts and Figures 2005 - Cancers linked to Infectious Diseases** [http://www.cancer.org/acs/groups/content/@research/documents/document/acspc-038821.pdf] Accessed: April 22, 2016
4. **Liver Cancer** [http://www.cancer.org/cancer/livercancer/detailedguide/liver-cancer-treating-by-stage] Accessed: April 23, 2016
5. **Primary liver cancer** [http://patient.info/health/primary-liver-cancer-leaflet] Accessed: April 23, 2016
6. **Hepatitis virus panel** [https://www.nlm.nih.gov/medlineplus/ency/article/003558.htm] Accessed: April 23, 2016
7. **Interpretation of Hepatitis B Serologic Test Results** [http://www.cdc.gov/hepatitis/HBV/PDFs/SerologicChartv8.pdf] Accessed: April 23, 2016
8. **Hepatitis C** [https://www.nlm.nih.gov/medlineplus/ency/article/000284.htm] Accessed: April 23, 2016
9. **Infections That Can Lead to Cancer** [http://www.cancer.org/cancer/cancercauses/othercarcinogens/infectiousagents/infectiousagentsandcancer/infectious-agents-and-cancer-viruses] Accessed: April 22, 2016
10. Kao JH: **Diagnosis of hepatitis B virus infection through serological and virological markers.** *Expert Rev Gastroenterol Hepatol* 2008, **2:**553-562.
11. **Viral Hepatitis** [https://www.atsu.edu/faculty/chamberlain/website/lectures/lecture/hepatit2.htm] Accessed: Aug 12, 2016
12. Datta S, Budhauliya R, Das B, Chatterjee S, Vanlalhmuaka, Veer V: **Next-generation sequencing in clinical virology: Discovery of new viruses.** *World J Virol* 2015, **4:**265-276.
13. Barzon L, Lavezzo E, Militello V, Toppo S, Palu G: **Applications of next-generation sequencing technologies to diagnostic virology.** *Int J Mol Sci* 2011, **12:**7861-7884.
14. Byron SA, Van Keuren-Jensen KR, Engelthaler DM, Carpten JD, Craig DW: **Translating RNA sequencing into clinical diagnostics: opportunities and challenges.** *Nat Rev Genet* 2016, **17:**257-271.
15. Wang F, Sun Y, Ruan J, Chen R, Chen X, Chen C, et al: **Using Small RNA Deep Sequencing Data to Detect Human Viruses.** *Biomed Res Int* 2016, **2016:**2596782.
16. Khoury JD, Tannir NM, Williams MD, Chen Y, Yao H, Zhang J, et al: **Landscape of DNA virus associations across human malignant cancers: analysis of 3,775 cases using RNA-Seq.** *J Virol* 2013, **87:**8916-8926.

17. Salyakina D, Tsinoremas NF: **Viral expression associated with gastrointestinal adenocarcinomas in TCGA high-throughput sequencing data.** *Hum Genomics* 2013, **7:**23.
18. **FASTQ Format** [https://en.wikipedia.org/wiki/FASTQ_format] Accessed: Dec 19, 2016
19. **NCBI FTP site for viruses** [ftp://ftp.ncbi.nlm.nih.gov/genomes/Viruses/] Accessed: April 24, 2016
20. Schelhorn SE, Fischer M, Tolosi L, Altmuller J, Nurnberg P, Pfister H, et al: **Sensitive detection of viral transcripts in human tumor transcriptomes.** *PLoS Comput Biol* 2013, **9:**e1003228.
21. Chen Y, Yao H, Thompson EJ, Tannir NM, Weinstein JN, Su X: **VirusSeq: software to identify viruses and their integration sites using next-generation sequencing of human cancer tissue.** *Bioinformatics* 2013, **29:**266-267.
22. Li JW, Wan R, Yu CS, Co NN, Wong N, Chan TF: **ViralFusionSeq: accurately discover viral integration events and reconstruct fusion transcripts at single-base resolution.** *Bioinformatics* 2013, **29:**649-651.
23. Wang Q, Jia P, Zhao Z: **VirusFinder: software for efficient and accurate detection of viruses and their integration sites in host genomes through next generation sequencing data.** *PloS one* 2013, **8:**e64465.
24. Kostic AD, Ojesina AI, Pedamallu CS, Jung J, Verhaak RG, Getz G, et al: **PathSeq: software to identify or discover microbes by deep sequencing of human tissue.** *Nature biotechnology* 2011, **29:**393-396.
25. Bhaduri A, Qu K, Lee CS, Ungewickell A, Khavari PA: **Rapid identification of non-human sequences in high-throughput sequencing datasets.** *Bioinformatics* 2012, **28:**1174-1175.
26. Lawrence M, Huber W, Pages H, Aboyoun P, Carlson M, Gentleman R, et al: **Software for computing and annotating genomic ranges.** *PLoS Comput Biol* 2013, **9:**e1003118.
27. **The Cancer Genome Atlas** [https://tcga-data.nci.nih.gov/] Accessed: July 14, 2016
28. **FASTA format** [https://en.wikipedia.org/wiki/FASTA_format] Accessed: April 24, 2016
29. Li B, Dewey CN: **RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.** *BMC Bioinformatics* 2011, **12:**323.
30. Bhuvaneshwar K, Sulakhe D, Gauba R, Rodriguez A, Madduri R, Dave U, et al: **A case study for cloud based high throughput analysis of NGS data using the globus genomics system.** *Comput Struct Biotechnol J* 2015, **13:**64-74.
31. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2009, **25:**1754-1760.
32. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2.** *Nat Methods* 2012, **9:**357-359.
33. **BAM** [http://genome.sph.umich.edu/wiki/BAM] Accessed: 2016 April 24
34. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25:**2078-2079.

35. **Picard** [ttp://broadinstitute.github.io/picard.] Accessed: 2016 April 24
36. **GFF/GTF File Format - Definition and supported options** [http://useast.ensembl.org/info/website/upload/gff.html] Accessed: Aug 12, 2016
37. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.** *Bioinformatics* 2010, **26:**139-140.
38. D. R Cox DO: *Analysis of Survival Data.* Chapman & Hall; 1984.
39. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et al: **From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline.** *Curr Protoc Bioinformatics* 2013, **43:**11 10 11-33.
40. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al: **VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing.** *Genome research* 2012, **22:**568-576.
41. Spencer DH, Tyagi M, Vallania F, Bredemeyer AJ, Pfeifer JD, Mitra RD, et al: **Performance of common analysis methods for detecting low-frequency single nucleotide variants in targeted next-generation sequence data.** *J Mol Diagn* 2014, **16:**75-88.
42. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ: **Second-generation PLINK: rising to the challenge of larger and richer datasets.** *Gigascience* 2015, **4:**7.
43. Echeverria N, Moratorio G, Cristina J, Moreno P: **Hepatitis C virus genetic variability and evolution.** *World J Hepatol* 2015, **7:**831-845.
44. Bonsall D, Ansari MA, Ip C, Trebes A, Brown A, Klenerman P, et al: **ve-SEQ: Robust, unbiased enrichment for streamlined detection and whole-genome sequencing of HCV and other highly diverse pathogens.** *F1000Res* 2015, **4:**1062.
45. In *Hepatitis C Viruses: Genomes and Molecular Biology.* Edited by Tan SL. Norfolk (UK); 2006
46. Stapleton JT, Schmidt WN, Katz L: **Seronegative hepatitis C virus infection, not just RNA detection.** *J Infect Dis* 2004, **190:**651-652.
47. Tang KW, Alaei-Mahabadi B, Samuelsson T, Lindh M, Larsson E: **The landscape of viral expression and host gene fusion and adaptation in human cancer.** *Nat Commun* 2013, **4:**2513.
48. Kremsdorf D, Soussan P, Paterlini-Brechot P, Brechot C: **Hepatitis B virus-related hepatocellular carcinoma: paradigms for viral-related human carcinogenesis.** *Oncogene* 2006, **25:**3823-3833.
49. Arbuthnot P, Kew M: **Hepatitis B virus and hepatocellular carcinoma.** *Int J Exp Pathol* 2001, **82:**77-100.
50. Nassal M: **HBV cccDNA: viral persistence reservoir and key obstacle for a cure of chronic hepatitis B.** *Gut* 2015, **64:**1972-1984.
51. Saksena NK, Wang B, Zhou L, Soedjono M, Ho YS, Conceicao V: **HIV reservoirs in vivo and new strategies for possible eradication of HIV from the reservoir sites.** *HIV AIDS (Auckl)* 2010, **2:**103-122.

52. Perkins RS, Sahm K, Marando C, Dickson-Witmer D, Pahnke GR, Mitchell M, et al: **Analysis of Epstein-Barr virus reservoirs in paired blood and breast cancer primary biopsy specimens by real time PCR.** *Breast Cancer Res* 2006, **8:**R70.
53. Felis-Giemza A, Olesinska M, Swierkocka K, Wiesik-Szewczyk E, Haladyj E: **Treatment of rheumatic diseases and hepatitis B virus coinfection.** *Rheumatol Int* 2015, **35:**385-392.
54. Gulley ML, Tang W: **Laboratory assays for Epstein-Barr virus-related disease.** *J Mol Diagn* 2008, **10:**279-292.
55. **Hepatitis C tests** [http://www.hepcnet.net/serologic_tests.html] Accessed: 2016 April 24
56. Liang TJ: **Hepatitis B: the virus and disease.** *Hepatology* 2009, **49:**S13-21.
57. Tage-Jensen U, Aldershvile J, Schlichting P: **Immunosuppressive treatment of HBsAg-positive chronic liver disease: significance of HBeAg.** *Hepatology* 1985, **5:**47-49.
58. Tran TT: **Immune tolerant hepatitis B: a clinical dilemma.** *Gastroenterol Hepatol (N Y)* 2011, **7:**511-516.
59. Park JH: **[Hepatitis B virus surface antigen: a multifaceted protein].** *Korean J Hepatol* 2004, **10:**248-259.
60. Locarnini S, Bowden S: **Hepatitis B surface antigen quantification: not what it seems on the surface.** *Hepatology* 2012, **56:**411-414.
61. Nelson PN, Carnegie PR, Martin J, Davari Ejtehadi H, Hooley P, Roden D, et al: **Demystified. Human endogenous retroviruses.** *Mol Pathol* 2003, **56:**11-18.
62. Tan Z, Pionek K, Unchwaniwala N, Maguire ML, Loeb DD, Zlotnick A: **The interface between hepatitis B virus capsid proteins affects self-assembly, pregenomic RNA packaging, and reverse transcription.** *J Virol* 2015, **89:**3275-3284.
63. Jain S, Chang TT, Chen S, Boldbaatar B, Clemens A, Lin SY, et al: **Comprehensive DNA methylation analysis of hepatitis B virus genome in infected liver tissues.** *Sci Rep* 2015, **5:**10478.
64. Cao GW: **Clinical relevance and public health significance of hepatitis B virus genomic variations.** *World J Gastroenterol* 2009, **15:**5761-5769.
65. Wang YZ, Zhu Z, Zhang HY, Zhu MZ, Xu X, Chen CH, et al: **Detection of hepatitis B virus A1762T/G1764A mutant by amplification refractory mutation system.** *Braz J Infect Dis* 2014, **18:**261-265.
66. Xiao L, Zhou B, Gao H, Ma S, Yang G, Xu M, et al: **Hepatitis B virus genotype B with G1896A and A1762T/G1764A mutations is associated with hepatitis B related acute-on-chronic liver failure.** *J Med Virol* 2011, **83:**1544-1550.
67. Robinson LR, Whelan SP: **Infectious Entry Pathway Mediated by the Human Endogenous Retrovirus K Envelope Protein.** *J Virol* 2016, **90:**3640-3649.
68. Boller K, Schonfeld K, Lischer S, Fischer N, Hoffmann A, Kurth R, et al: **Human endogenous retrovirus HERV-K113 is capable of producing intact viral particles.** *J Gen Virol* 2008, **89:**567-572.

69. Moyes DL, Martin A, Sawcer S, Temperton N, Worthington J, Griffiths DJ, et al: **The distribution of the endogenous retroviruses HERV-K113 and HERV-K115 in health and disease.** *Genomics* 2005, **86:**337-341.

70. Downey RF, Sullivan FJ, Wang-Johanning F, Ambs S, Giles FJ, Glynn SA: **Human endogenous retrovirus K and cancer: Innocent bystander or tumorigenic accomplice?** *Int J Cancer* 2015, **137:**1249-1257.

FIGURES, TABLES AND ADDITIONAL FILES

Table 1. **Comparison of existing pipelines that detect viruses from human transcriptome data**

Table 2A (top) and 2B (bottom): Comparison of viral detection from serology (blood) with viral detection from RNA-seq data (tumor tissue) for Hepatitis B and Hepatitis C respectively. The results shown are from Module 1 ('Filtered human sample input') performed using Bowtie2 alignment tool

Table 3: Summary of genome browser pattern for Hepatitis B genome with Dead/Alive status

Table 4: Differential expression analysis of transcript level read counts on Hep B sub-group comparing Dead and Alive samples. These results shown used the viral-gene/CDS data obtained from Module 1 (using alignment tool Bowtie2) + Module 3. The table shows results with q value < 0.06 and sorted based on LogFC in the descending order.

Table 5: Cox proportional hazard survival analysis (across 25 HepB samples and 25 HepB + HepC Samples). These results shown used the viral-gene/CDS data obtained from Module 1 (using alignment tool Bowtie2) + Module 3. *Coef: coefficient (Beta) of the model; exp(coef): Hazard Ratio; se(coef) : Standard Error; Pr(>|z|) : P-value*

Table 6: Results of case-control association test applied on the results from viral variant calling (showing only common results between two possible analysis steps). The table is sorted based on Annotation. Annotation includes gene name, protein name, etc., separated by commas, multiple annotations separated by semi-colon

Figure 1: viGEN pipeline. Each module has a color, shown in the legend

Figure 2: Landscape of HBV genome across the 25 HepB samples. The BAM files used were from Module 1 generated using Bowtie2. The image shows the three types of patterns of pileup distribution namely full (red dot), truncated (blue dot), and empty (black dot).

Additional File 1: viGEN Github tutorial

Additional File 2: List of viruses that had genome read count more than 100 and short listed for analysis in Module 3 and 4

Additional File 3: Landscape of viruses at the genome level

Additional File 4: Results of case-control association test applied on the viral variant calling results