G&I
GENOMICS & INFORMATICS

Original article

eISSN 2234-0742
Genomics Inform 2022; 20(4):e44
https://doi.org/10.5808/gi.22047

Received: July 26, 2022
Revised: September 3, 2022
Accepted: September 25, 2022

*Corresponding author:
E-mail: Chureerat.P@chula.ac.th

**Corresponding author:
E-mail: sp.medbiochemcu@gmail.com

Metagenomic analysis of viral genes integrated in whole genome sequencing data of Thai patients with Brugada syndrome

Suwalak Chitcharoen[1,2], Chureerat Phokaew[3,4,5]*, John Mauleekoonphairoj[6,7], Apichai Khongphatthanayothin[6,8,9], Boosamas Sutjaporn[4,6], Pharawee Wandee[6], Yong Poovorawan[10], Koonlawee Nademanee[6,11,12], Sunchai Payungporn[2]**

[1]Program in Bioinformatics and Computational Biology, Graduate School, Chulalongkorn University, Bangkok 10330, Thailand
[2]Center of Excellence in Systems Microbiology, Department of Biochemistry, Faculty of Medicine, Chulalongkorn University, Bangkok 10330, Thailand
[3]Center of Excellence for Medical Genomics, Medical Genomics Cluster, Faculty of Medicine, Chulalongkorn University, Bangkok 10330, Thailand
[4]Excellence Center for Genomics and Precision Medicine, King Chulalongkorn Memorial Hospital, The Thai Red Cross Society, Bangkok 10330, Thailand
[5]Research Affairs, Faculty of Medicine, Chulalongkorn University, Bangkok 10330, Thailand
[6]Department of Medicine, Faculty of Medicine, Center of Excellence in Arrhythmia Research Chulalongkorn University, Chulalongkorn University, Bangkok 10330, Thailand
[7]Interdisciplinary Program of Biomedical Sciences, Graduate School, Chulalongkorn University, Bangkok 10330, Thailand
[8]Division of Cardiology, Department of Pediatrics, Faculty of Medicine, Chulalongkorn University, Bangkok 10330, Thailand
[9]Bangkok General Hospital, Bangkok 10330, Thailand
[10]Department of Pediatrics, Faculty of Medicine, Chulalongkorn University, Bangkok 10330, Thailand
[11]Department of Medicine, Faculty of Medicine, Chulalongkorn University, Bangkok 10330, Thailand
[12]Pacific Rim Electrophysiology Research Institute, Bumrungrad Hospital, Bangkok 10110, Thailand

Brugada syndrome (BS) is an autosomal dominant inheritance cardiac arrhythmia disorder associated with sudden death in young adults. Thailand has the highest prevalence of BS worldwide, and over 60% of patients with BS still have unclear disease etiology. Here, we performed a new viral metagenome analysis pipeline called VIRIN and validated it with whole genome sequencing (WGS) data of HeLa cell lines and hepatocellular carcinoma. Then the VIRIN pipeline was applied to identify viral integration positions from unmapped WGS data of Thai males, including 100 BS patients (case) and 100 controls. Even though the sample preparation had no viral enrichment step, we can identify several virus genes from our analysis pipeline. The predominance of human endogenous retrovirus K (HERV-K) viruses was found in both cases and controls by blastn and blastx analysis. This study is the first report on the full-length HERV-K assembled genomes in the Thai population. Furthermore, the HERV-K integration breakpoint positions were validated and compared between the case and control datasets. Interestingly, Brugada cases contained HERV-K integration breakpoints at promoters five times more often than controls. Overall, the highlight of this study is the BS-specific HERV-K breakpoint positions that were found at the gene coding region "*NBPF11*" (n = 9), "*NBPF12*" (n = 8) and long non-coding RNA (lncRNA) "*PCAT14*" (n = 4) region. The genes and the lncRNA have been reported to be associated with congenital heart and arterial diseases. These findings provide another aspect of the BS etiology associated with viral genome integrations within the human genome.

Keywords: Brugada syndrome, human endogenous retrovirus K, metagenome, VIRIN, virus integration breakpoint, whole genome sequencing

Introduction

Brugada syndrome is an inherited arrhythmogenic disease leading to a high risk of acute cardiac death. This syndrome has been connected to a genetic variant with an autosomal dominant inheritance pattern [1,2]. The highest prevalence is in Southeast Asia, especially in Thailand (6.8 per 1,000), where it is almost 15 times higher than worldwide [2-4]. Hundreds of gene variants have been associated with Brugada syndrome, of which the mutation in *SCN5A* or *SCN10A* genes of sodium channels has been commonly found with the disease (30%) [5,6]. While new findings, the *MAPRE2* mechanism, and the microtubule-related trafficking effects on NaV1.5 expression have been explored in Brugada syndrome by genome-wide association analysis [5], for almost 60% of patients, etiologic causes are still unknown [7].

The whole genome sequencing (WGS) data have become highly valuable information that can be used to screen all variants, allele assignment, insertion, and detection of structural variation [8,9]. Generally, after mapping WGS reads to the human reference genome, there remains 5%–10% unmapped reads [10]. The unmapped reads may contain microbial agents, especially viral elements, due to the integrative capacity of various viruses [11]. The metagenomics approach is suitable for identifying uncharacterized sequencing reads [12].

Approximately 8% of the human genome contains human endogenous retroviruses. The human endogenous retrovirus (HERV)'s transcripts and regulatory functions have been identified in numerous diseases [13,14] including multiple sclerosis [15], diabetes [16], systemic lupus erythematosus [17], psoriasis [18], rheumatoid arthritis [19], and cancer [20]. Moreover, human papillomavirus (HPV), hepatitis B virus (HBV), and Epstein-Barr virus (EBV) are exogenous viruses that are associated with diseases. They are well-known as insertion viruses commonly found in the human genome and can induce tumorigenesis and cancer (10%–15% of all cancer) [21,22]. Local viral integrations may cause genomic instability followed by altered gene copy numbers and gene expression around the integration sites. Therefore, these inserted positions provide valuable information for understanding the mechanisms of virus-related diseases and the etiologic [23].

Additionally, the infections such as enteroviruses (coxsackievirus, enterovirus, echovirus) and adenoviruses play an important role in sudden cardiac death [24,25]. The most common cardiotropic viruses are EBV, coxsackievirus, adenovirus, human herpesvirus 6 (HHV6), cytomegalovirus, hepatitis C virus, and parvovirus B19. Moreover, parvovirus B19 was also associated with Brugada syndrome. They potentially trigger an autoimmune response against components of the heart or mediate direct cardiac injury [26]. Thus, viral genes and integration positions in Brugada syndrome patients are useful evidence that can be used to discover the disease's etiology and progression.

We identified the putative viral gene and protein in 200 Thai male WGS data. We further developed an analysis pipeline to identify virus integration positions in human genome sequencing data called "VIRIN". The human endogenous retrovirus K (HERV-K) genomes were assembled and explored in two potential integration loci of HERV-K, namely the "*neuroblastoma breakpoint genes family* (*NBPF*)" gene family and long non-coding RNA (lncRNA) "*PCAT14*", which are related to Brugada syndrome from WGS data of an individual Thai patient with Brugada syndrome and a control volunteer.

Methods

Study cohort

The study cohort was divided into two groups, the Brugada cases and controls. The cases consisted of 100 Thai male subjects with type I Brugada electrocardiogram (ECG) using the criteria of the 2013 Heart Rhythm Society/European Heart Rhythm Association/Asia Pacific Heart Rhythm Society Expert Consensus Statement. The training physicians read and confirmed all ECGs of cases with a type I Brugada pattern. The 100 Thai male control subjects comprised those who had a standard 12-lead ECG without type I Brugada. Both groups were representative of a wide age range (between 19 and 75 years, with medians of 50 years in the case group and 47 years in the control group). All subjects were of Thai ethnicity by self-report. The Ethics Review Committee of all the institutions approved the study (NCT04232787).

All blood samples were collected, and DNA was extracted as described in a previous report. The sequencing libraries were prepared with a polymerase chain reaction–free reaction [9]. Then, the human genomic DNA libraries were sequenced by Illumina HiseqX platforms (Cambridge, UK) with a pair-end sequencing (2 × 150 bp) strategy [9].

Extraction of the unmapped and soft-clipped sequence

First, the raw FASTQ reads were filtered with Trimmomatic version 0.38 [27] by sliding window at mapping quality 30; all reads shorter than 50 nucleotides were removed. Second, the filtered reads were mapped to the NCBI Genome Reference Consortium Human Build 38 (GRCh38) with decoys reference using iSAAC-03.16.02.19 (version 0.7.16a) with a default setting. An unmapped read whose mate is mapped was extracted using the

SAMtools 1.15 "view -f 4 -F 264" command [28]. Furthermore, the soft-clipped reads with GRCh38 were extracted using the modified extractSoftclipped function of the SE-MEI tools (https://github.com/dpryan79/SE-MEI).

Identification of the viral sequences

The viral sequence identification was performed on the NCBI 13,434 complete reference viral database (https://ftp.ncbi.nlm.nih.gov/refseq/release/viral/, downloaded on January 7, 2021) by the Bowtie2 aligner version 2.3.5.1 [29]. Next, the identified viral sequence was merged by *de novo* assembly approach utilizing SPAdes v3.13.0 with the "-k 33" option [30]. Then, to identify the viral integration site, the contigs were validated with the blastn in NCBI BLAST+ [31] and Diamond blastx v2.0.15.153 [32], respectively.

The original extracted single-mate unmapped and soft-clipped reads were re-aligned with the selected virus reference genome. The reads mapped with the virus reference genome were extracted from the original GRCh38 human alignment bam file. The position of the virus integration was reported by the in-house bash script and BEDtools v2.27.1 command [33]. The virus integration positions (breakpoints) were annotated with DNase I hypersensitivity regions, Repeatmasker, and GencodeV.40 [34-36]. The whole sequence of steps of the analytical source code is available on GitHub (https:// gist.github.com/Suwalak-Chit/VIRIN).

Data interpterion and statistical analysis

Statistical analysis and visualization were performed using GraphPad Prism 8.0.1 software. The descriptive statistics and continuous variables consistent with a normal distribution were represented by means and standard deviations; non-parametric tests were performed with t-tests or the Mann-Whitney U test. $p < 0.05$ was considered statistically significant.

Data availability

The data in this study were available from the National Research Council of Thailand under license for the current study and were not publicly available. The data were available from the authors upon a reasonable request and with the permission of the National Research Council of Thailand.

Results

The characterization of unmapped reads

The WGS data from both case (n = 100) and control (n = 100) datasets are 929,286,470 and 935,775,065 pair-end read, respectively. Most reads were aligned with the human reference genome 96.48% (case) and 96.62% (control). The average unmapped reads which remained amounted to 44,548,377 reads (case) and 44,784,395 reads (control). Among the mapped reads, the soft-clipped reads were 9,009,042 in the case group and 8,860,802 in the control group (Fig. 1). The number of unmapped reads between groups did not significantly differ in the t-test ($p = 0.172$).

Viral gene or/and genome identification

The unmapped and soft-clipped reads were aligned against the NCBI viral genome using Bowtie2. With a cut-off of at least 1,000 reads containing 90 viruses in each sample, 291,126,786 reads were assigned to the 285 viral references. The SPAdes assembly tools [37] were used for *de novo* assembly of the virus mappable read in each sample. The whole contigs were hit with eight virus genomes (the contig length >5% of each virus genome) by Blastn. Three viruses (Torque teno virus 10 [TTV 10], human endogenous retrovirus K, and Bat associated circovirus 4) were found in both datasets. Five viruses (Torque teno virus 19, Torque teno virus 3, Torque teno virus 8, Aeribacillus virus AP45, and Gemykibivirus humas3) were found only in the case dataset. Seven of the eight identified are DNA viruses (Supplementary Table 1). Interestingly, the HERV-K and Bat-associated circovirus four genome completeness significantly differed between the cases and controls ($p < 0.001$) (Fig. 2). Almost full-length genome of human endogenous retrovirus K was observed for the datasets (the largest contig in case, 86.54%; control, 84.74% genome completeness) (Fig. 2).

The translated nucleotide blast on protein database (blastx) result showed 13 viral proteins positive in more than ten samples in each group. The human endogenous retrovirus K putative envelope protein was detected in all subjects from the blastx results (Supplementary Table 2). The HERV-K genome was assembled and the genome coverage was visualized. For the case subject data, the genome depth coverages on each HERV-K gene were gag gene (194.44x), pro gene (410.79x), pol gene (203.53x), and env gene (189.76x), respectively. In contrast, the genome coverage of HERV-K genes in the control subject data included the gag gene (199.44x), pro gene (610.79x), pol gene (243.53x), and env gene (129.76x) (Fig. 3). We also detected the A55 protein of the BeAn virus genome in 61 cases and 70 control data. Moreover, 11 form 13 (84.61%) of the identified viral proteins were retrovirus proteins.

VIRIN analysis pipeline validation

The viral breakpoint integration position was identified with the VIRIN analysis pipeline. The pipeline was validated with the cervical cancer cells (HeLa cells) and hepatocellular carcinoma (HCC) tissue WGS data, in which well-known HPV-18 and HBV

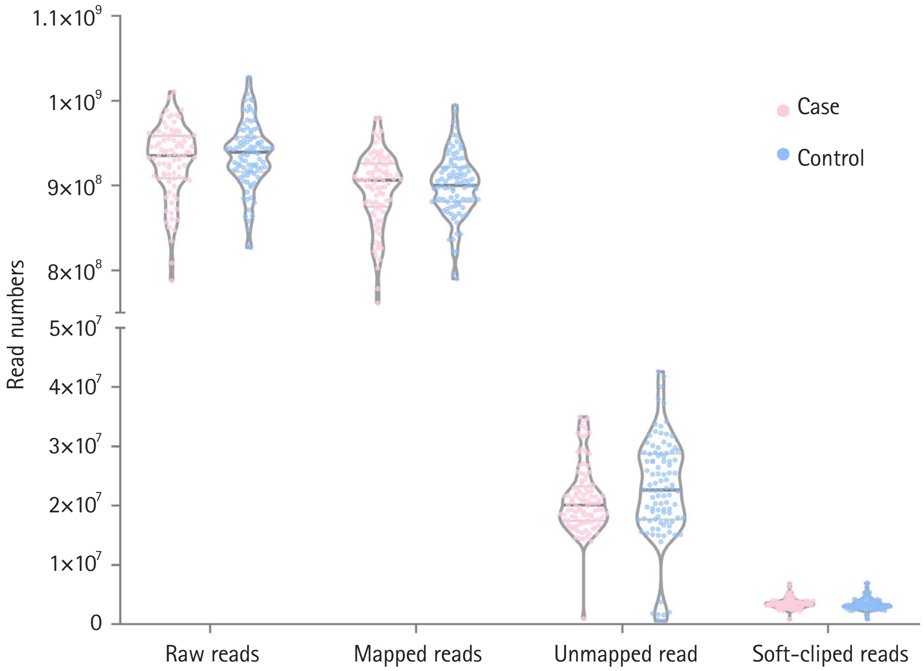

Fig. 1. Distribution number of mapped and unmapped reads with GRCh38.

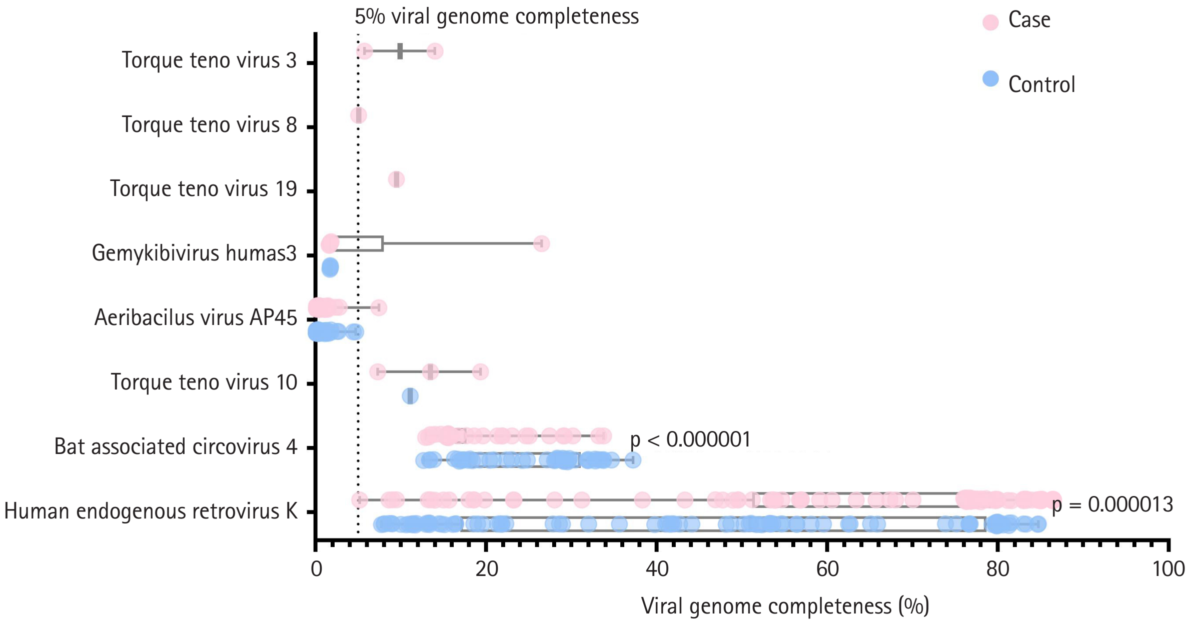

Fig. 2. The longest contigs size of each identified virus by nucleotide blast from individual sample.

are integrated into the cellular genome, respectively [38]. The result from VIRIN showed HPV-18 was integrated into the three breakpoints of locus 8q24.21 (Chr8:127,222,011, Chr8:127, 218,387 and Chr8: 127,229,303) of HeLa cell WGS data (SRR5009881) [39]. Moreover, the analysis pipeline identified two HBV breakpoints at locus 17p13.1 (Chr17:10,110,360 and Chr17:10,366,141) from HCC tissue WGS data (ERR173408 and ERR181167) [40] (Supplementary Table 3). This evidence suggests that VIRIN is effective for the identification of virus integration breakpoints.

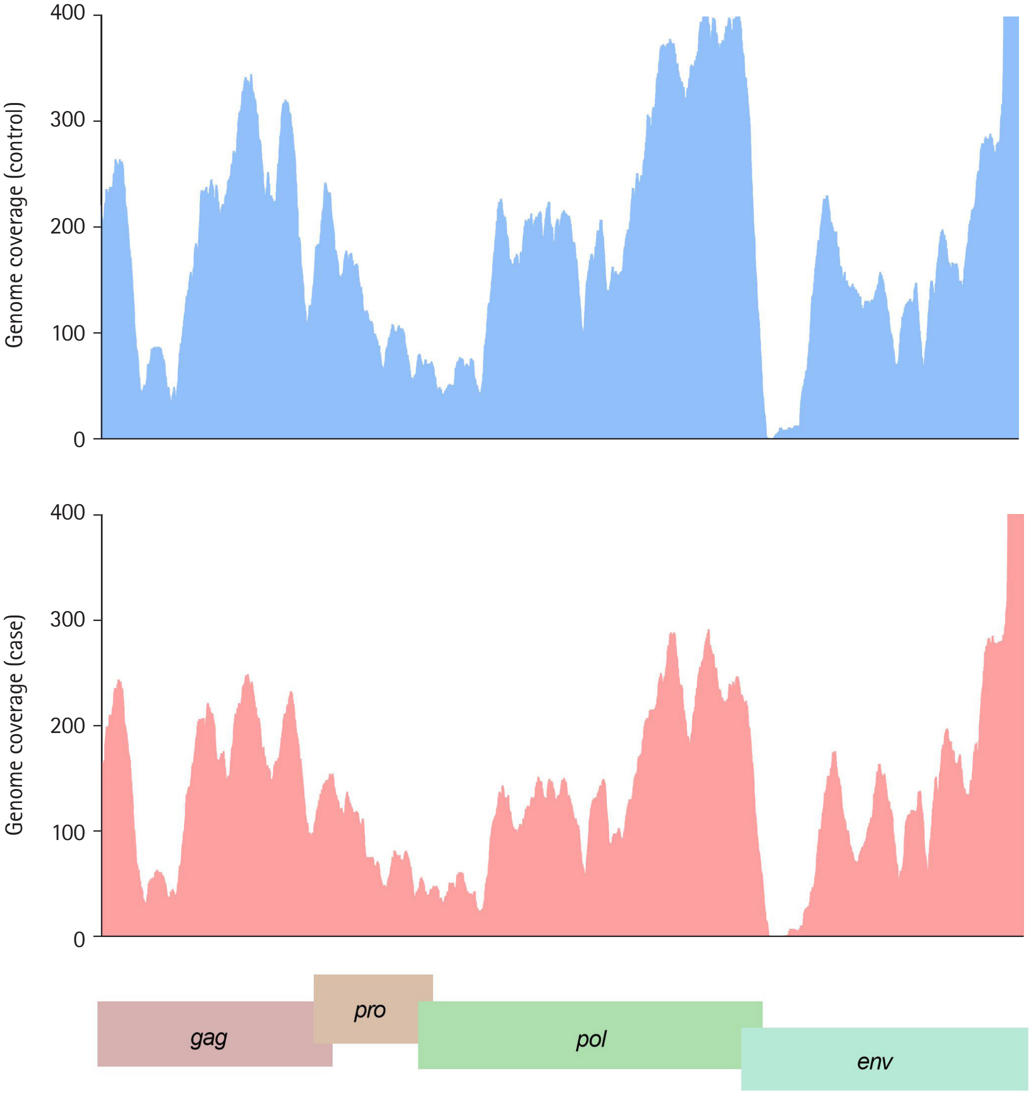

Fig. 3. The genome depth coverage of human endogenous retrovirus K (HERV-K) in representative case and control.

HERV-K breakpoint position identification

We validated the HERV-K breakpoints in the human genome. The total breakpoints were 952 in the case group and 814 in the control group datasets. The number of breakpoints was significantly different between the two datasets (Mann-Whitney U test, p = 0.0028) (Fig. 4A). Chromosome 1 had the highest HERV-K breakpoints in both datasets (case, 38; control, 20), while chromosomes 13, 15, and 20 had no HERV-K breakpoint position. Chromosomes 2, 12, 16, and X contained the HERV-K breakpoints only in the case dataset (4, 3, 8, and 7, respectively). However, there were no significant differences (t-test, $p < 0.05$) in HERV-K insertion in any chromosomes between the case and control (Supplementary Table 4). Moreover, the number of HERV-K breakpoints in DNase I hypersensitivity regions were significantly different between case group (81 breakpoints) and control group (40 breakpoints) (Mann-Whitney U test, p = 0.012) (Fig. 4B). More than 60% of HERV-K breakpoints were located in the intergenic region (case, 63.89%; control, 69.15%). The breakpoint in case datasets was more often in promoter than control datasets (15.28% and 1.06%, respectively) (Fig. 4C). The HERV-K breakpoints at the repeat region include 54 and 26 breakpoints from 30 cases and 20 controls, respectively. HERV-K breakpoints in the case group (32 breakpoints) were located mostly on the long terminal repeat (LTR), while most HERV-K breakpoints from the control group (12 breakpoints) were located on the long interspersed nuclear element (LINE) (Table 1).

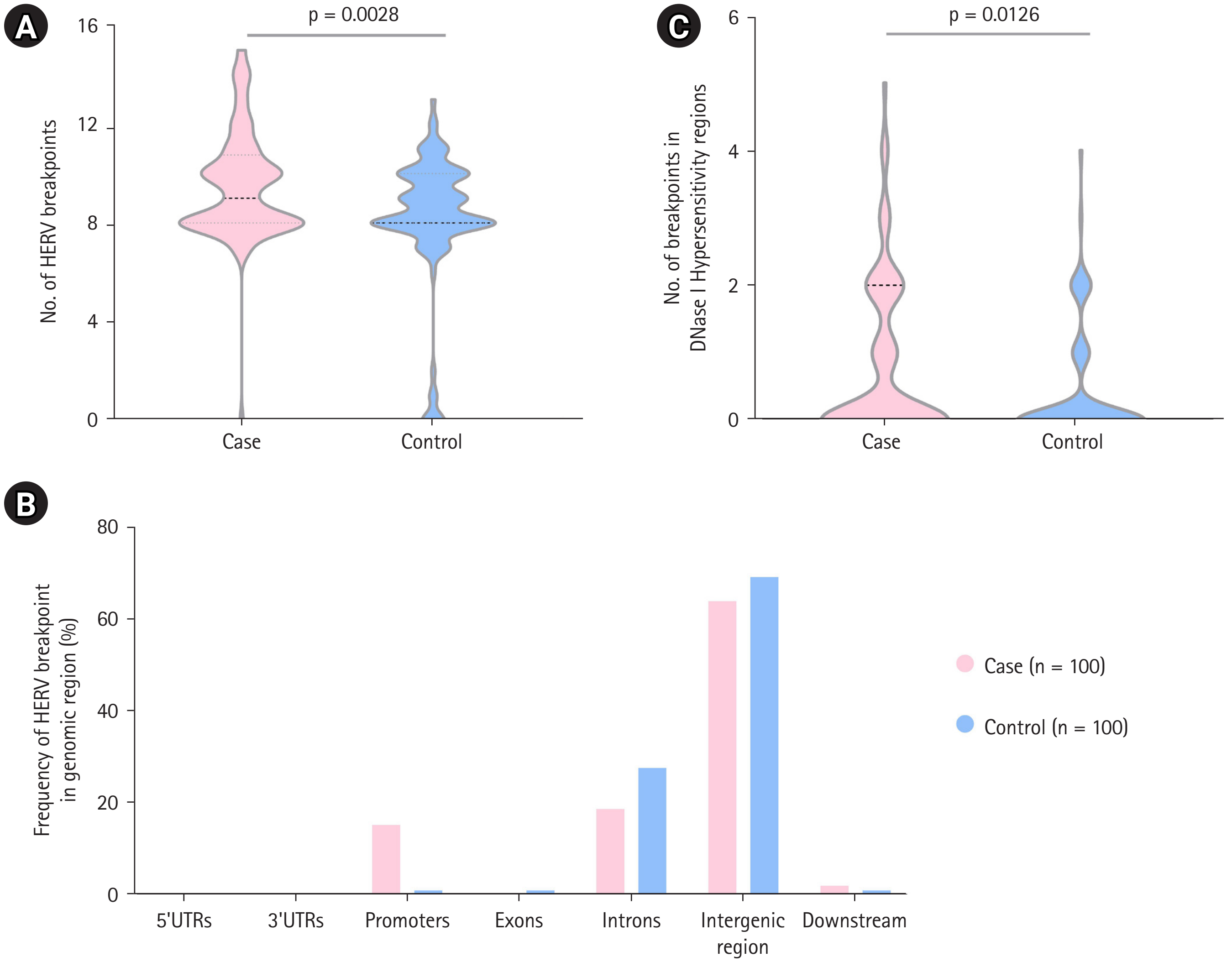

Fig. 4. The number of human endogenous retrovirus K (HERV-K) breakpoints in the human genome individual case (n = 100) and control group (n = 100) HERV-K breakpoints (A), the total distribution of HERV-K breakpoints in the genomic region (B), and the individual HERV-K breakpoints in DNase I hyposensitivity regions (C). UTR, untranslated region.

Table 1. The number of breakpoints in four types of repeat regions from each dataset

Repeat region	Case (n = 32)	Control (n = 20)
LTR	32	6
LINE	15	12
Simple repeat	3	0
Retroposon	3	7
SINE	1	1
Total	54	26

LTR, long terminal repeat; LINE, long interspersed nuclear element; SINE, short interspersed nuclear element.

All HERV-K breakpoints were annotated on the gene (gencodeV.40), including 27 positions from 27 cases and 29 positions from 18 controls. The HERV-K breakpoints were located in protein-coding (case, 17; control, 20), lncRNA (case, 8; control, 7) and a small number of pseudogene (case, 2; control, 3) (Supplementary Table 5). An individual unique HERV-K breakpoint was also found in 16 cases and 21 controls. Importantly, 25 HERV-K breakpoints from nine Brugada syndrome cases were located in a *NBPF*. Additionally, the HERV-K breakpoint from four Brugada syndrome samples was located on the lncRNA name *PCAT14* (prostate cancer-associated transcript 14). Even though the HERV-K breakpoint in the nicotinamide nucleotide transhydrogenase function (*NNT*) gene was found in many samples, more were detected in the control dataset (n = 10) than in the case dataset (n = 5). Meanwhile the *NBPF11*, *NBPF12*, and *PCAT14* were found only in the case group, and not detected in the control dataset (Tables 2 and 3).

Discussion

Viral gene/genome detection

A viral metagenome analysis was performed on the Brugada syndrome cases and control WGS data, to investigate the viral genome integration from the WGS unmapped read in each sample. Our data analysis found the genes of 90 integrated viruses from both human and non-human viruses. The human viruses in this study were members of the *Herpesviridae* family and HERV, the same as in previous reports [12,41]. Almost 80% of virus-aligned reads were assigned as dsDNA virus because the library preparation kit was appropriated for DNA source, which similar to the result of virome in human WGS from the previous study [42]. According to the library preparation approach, giant viruses (large genome viruses) such as the Pandoravirus (2 Mbp genome size) were observed in our study as well as in the previous report [43].

The *de novo* assembled contigs were aligned with the virus reference genomes (NCBI). Eight viruses were kept after excluding the virus contigs that contained less than 5% of their genome. TTV, a dsDNA virus belonging to the *Anelloviridae* family, had considerable genetic variability and extreme diversity [44]. In previous reports, the TTV DNA was detected in secretions of healthy humans, such as blood, saliva, breast milk, tears, bile, and urine [45,46]. The TTVs DNA level was also considered the marker of the immunological status, hepatitis, gastroenteritis, periodontitis, multiple sclerosis, and cancer [47,48]. In this study, we have found a unique TTV19, TTV8 in a specific case, TTV3 in two cases, and TTV10 in both groups (case, 3; control, 1). Nevertheless, several virome studies of human blood frequently found TTV in the sam-

Table 2. Gene annotation list in breakpoints of HERV-K integration in case dataset

No.	Gene name	Gene type	Chr	Position	Frequency
1	*NBPF12*	Protein coding	1	146948791 146949149 146949314	8
2	*NBPF11*	Protein coding	1	148137980	9
3	*SIL1*	Protein coding	5	139112931	1
4	*NNT*	Protein coding	5	43665587 43665608 43665610 43665611	5
5	*PDE10A*	Protein coding	6	165500565	1
6	*SSBP1*	Protein coding	7	141752231	1
7	*TCP11L1*	Protein coding	11	33049984	1
8	*ASRGL1*	Protein coding	11	62378218 62378902	2
9	*PPTC7*	Protein coding	12	110571019	1
10	*ZNF140*	Protein coding	12	133093110	1
11	*MAPK1*	Protein coding	22	21848137	1
12	*TMEM51-AS1*	lncRNA	1	15135632	1
13	*NEPRO-AS1*	lncRNA	3	113025862	1
14	*LINC02614*	lncRNA	3	125896808	1
15	*ENSG00000272462*	lncRNA	6	25999489	1
16	*ENSG00000285784*	lncRNA	9	11895136	1
17	*PCAT14*	lncRNA	22	23541467 23541739 23541740	4
18	*PDCL3P4*	Pseudogene	3	101695806	4
19	*ENPP7P15*	Pseudogene	11	3451290	1

HERV-K, human endogenous retrovirus K.

Table 3. Gene annotation list in breakpoints of HERV-K integration in control dataset

No.	Gene name	Gene type	Chr	Position	Frequency
1	*PRDX1*	Protein coding	1	45514062	1
2	*IPP*	Protein coding	1	45739955	1
3	*MPZL1*	Protein coding	1	167771343	1
4	*CR1*	Protein coding	1	207636105	1
5	*CR1*	Protein coding	1	207637031	1
6	*NNT*	Protein coding	5	43665299 43665592 43665597 43665598 43665601 43665607 43665608 43665612 43665614	10
7	*IQGAP2*	Protein coding	5	76550644	1
8	*ARMT1*	Protein coding	6	151456042	1
9	*LHFPL3*	Protein coding	7	104750438	1
10	*ABCC2*	Protein coding	10	99827666	1
11	*TTC5*	Protein coding	14	20269289	1
12	*SLC47A1*	Protein coding	17	19505358	1
13	*ENSG00000285988*	lncRNA	10	6830418	1
14	*ENSG00000255947*	lncRNA	11	61655408	1
15	*ENSG00000259048*	lncRNA	14	38118365	1
16	*ENSG00000287879*	lncRNA	18	79960352	1
17	*ENSG00000283907*	lncRNA	19	35575243	1
18	*ENSG00000286667*	lncRNA	19	386213	1
19	*MIR548XHG*	lncRNA	21	18568413	1
20	*SEC22B4P*	Pseudogene	1	146381378	1
21	*PDCL3P4*	Pseudogene	3	101693108	1

HERV-K, human endogenous retrovirus K; lncRNA, long non-coding RNA.

ples [49], while we reported only eight samples in total. Indeed, we found TTV reads in a total of 20 samples; however, some TTV contigs' number, coverage and length are too low. Since the human WGS library preparation has a virus enrichment step, we found TTV in a few samples. However, according to our findings, we could not confirm that TTV integration could be related to Brugada syndrome. On the other hand, the non-human viruses were known as the contaminant DNA from the reagents or sequencing process, including bacteriophage (Aeribacillus virus AP45) and mammalian virus (Bat associated circovirus four and Gemykibivirus humas3) [50]. The previously frequently found viruses associated with a cardiac infection, such as HHV6, EBV, and hepatitis C virus were also detected but could not be assembled to the large contigs [26]. Even though Parvoviruses B19 was reported as being associated with Brugada syndrome, it was not found in our data.

VIRIN validation

Previous research has reported that three virus segments of HPV18 are integrated into the HeLa genome on chromosome 8 (locus 8q23-24) upstream of the *myc* gene [51]. The result of VIRIN analysis pipelines was found in three HPV-18 integrated positions in the locus 8q23.21 of HeLa cell line WGS data. Generally, the HBV integration breakpoints in HCC are various. Most HBV breakpoints are near coding genes, including the *TERT*, *MLL4*, *CCNE1*, *SENP5*, and *ROCK1*. Recurrent HBV breakpoints occur within or close to repetitive, non-coding sequences, such as LINEs, Alu short interspersed elements, and LTR [52]. Our result also showed HBV integrated into the coding region of the *MYH13* gene at 17p13.1 in HCC tissue WGS data, similar to a previous report [53]. Thus, the VIRIN analysis pipeline can identify virus integration in human WGS data.

HERV-K breakpoint positions

In addition, we can assemble the full-length HERV-K genome in all datasets. The translated nucleotide blast on protein databases (blastx) result showed that the envelope proteins (env) of HERV-K were found in all datasets. In our data analysis, we also detected known DNA proviruses, including gag, pro, pol, and env, of HERV-K [54]. Since we utilized the WGS data for the blast, our data were not a direct viral protein identification. This is the first study demonstrating the HERV-K breakpoints in the WGS data of the Thai population. The polymorphic integration position of HERV-K could influence both viral protein production and host gene regulation. The specific HERV-K breakpoints might be associated with the potential pathogenicity in different individuals, for example, neurologic and immunologic diseases [54].

DNase I hypersensitivity regions are the important genomic landmarks for functionally active open chromatin. A previous report also showed that 15% of HERVs inserts are in the DNase I hypersensitivity region [55-57]. Interestingly, our results showed higher HERV-K breakpoints at DNase I hypersensitivity regions in the case group compared to the control group. Moreover, the breakpoints of HERV-K were located at the promoter regions 15 times more often in the case group compared to the control group, and this might be linked to the gene regulation process of Brugada syndrome pathogenesis. Moreover, Brugada syndrome is 8-10 times more prevalent in men than women [58], and our result showed HERV-K was integrated at chrX only in the case group (7 cases and 0 control). The previous studies also found a variant on the *KCNJ5* (potassium voltage-gated channel subfamily E regulatory subunit 5) gene located on chromosome X in Japanese patients with Brugada syndrome [59,60]. However, our result did not find HERV-K breakpoints in any gene regions. The potential of sex hormones for cardiac regulation is through the ion channel because the cardiac muscle found the main gonadal steroid receptors. Moreover, many ion channels, such as *CACNA1C* and *SCN5A*, are very sensitive to testosterone, and this could explain the gender difference in the prevalence of Brugada syndrome [61,62]. The highest number of HERV-K breakpoints is in the LTR from all the repeated regions in both groups. Although the HERV-K were generally integrated into the LTR regions, the number of the HERV-K breakpoints in the case group LTR region was five times higher than in the control group [63].

Importantly, our data found that the HERV-K breakpoints were in the protein-coding and lncRNA region. HERV-K breakpoints within the *NNT* gene were found in many samples (control, 10; case, 5). NNT is a proton pump in the inner mitochondrial membrane found throughout the human body. It is highly expressed in the adrenals, bladder, kidneys, thyroid, adipose tissue, and especially in the heart [64,65]. Several reports showed that the lack of the *NNT* gene triggers the down-regulation of glucocorticoid levels, inhibiting cardiovascular conditions. This finding is linked to our study that found more HERV-K breakpoints in the *NNT* gene in the control group than in the Brugada syndrome patient group [66,67]. However, the gene expression level of NNT protein in the heart plays an important role in the pathogenesis; thus, NNT protein levels in Brugada syndrome are interesting and need further investigation.

Most of the HERV-K breakpoints in the gene region were the *NBPF* genes, which were implicated in several developmental and neurogenetic diseases and congenital heart disease [17,68]. Furthermore, the NBPF family, such as NBPF1 and NBPF11, were reported as the translocation disrupts a sodium channels gene on chromosome 17 called *ACCN1* (amiloride-sensitive cation channel 1) [69,70]. Even if the mechanism of the *NBPF* gene family to electrophysiology is still unknown, it is possible that the HERV-K breakpoints on *NBPF* genes are related to Brugada syndrome pathogenesis. However, the underlying mechanism needs to be further investigated.

Several lncRNAs play an important role in cardiovascular diseases, such as lncRNA *LIPCAR* dysreglulation in heart failure and lncRNA *MIAT* upregulations in myocardial infarction [71-73]. A lncRNA gene, *PCAT14*, plays an important role in tumorigenesis in HCC and prostate cancer [74,75]. *PCAT14* expression is also an important prognostic for predicting metastatic disease. Furthermore, *PCAT14* contains the single nucleotide variants of SNP rs73155085-A and rs131408-C. The rs73155085-A and rs131408-C have been reported to be associated with coronary artery disease and peripheral arterial disease, respectively [17,76]. Thus, the breakpoint at the *PCAT14* gene is potentially involved with the Brugada syndrome pathogenesis. Moreover, *PCAT14* has been associated with the hormone testosterone in prostate cancer [77].

In conclusion, some key findings have emerged from this work. The HERV-K genome and their breakpoints in the Thai population genome have been reported, and the HERV-K breakpoint positions have been found in the data. Two (*NBPF* gene and *PCAT14* lncRNA) of these breakpoints have a reasonable potential to be key pathogenesis features of Brugada syndrome. Hence, these findings provide a new viewpoint on the etiology of Brugada syndrome, including the association with viruses and virus integration positions, and not limited to purely human genetics.

ORCID

Suwalak Chitcharoen: https://orcid.org/0000-0002-9110-6884
Chureerat Phokaew: https://orcid.org/0000-0002-4246-2604
John Mauleekoonphairoj: https://orcid.org/0000-0001-7644-0940
Apichai Khongphatthanayothin: https://orcid.org/0000-0001-5738-1738
Boosamas Sutjaporn: https://orcid.org/0000-0003-3117-9288
Pharawee Wandee: https://orcid.org/0000-0002-2676-0767
Yong Poovorawan: https://orcid.org/0000-0002-2337-6807
Koonlawee Nademanee: https://orcid.org/0000-0001-8139-0343
Sunchai Payungporn: https://orcid.org/0000-0003-2668-110X

Authors' Contribution

Conceptualization: SC, CP, SP. Data curation: SC, CP, JM. Formal analysis: SC. Funding acquisition: SC, CP, AK, YP, KN, SP. Methodology:SC, CP, JM, BS, PW, SP. Writing - original draft: SC. Writing - review & editing: CP, JM, AK, YP, SP.

Conflicts of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgments

We acknowledge CMKL university and PMU-C for providing computational resources use in storage and analysis of this study. We thank members of the Brugada consortium and the Thai Red Cross Society and Centre of Excellence in Clinical Virology, Faculty of Medicine, Chulalongkorn University for assistance with data collection and sequencing. We also would like to thank the Developer and Promotion Talent (DPST) Thailand scholarship Fund for Ph.D. student.

This research project is supported by the National Research Council of Thailand grand and the National Science and Technology Development Agency (P-15-5004) (to Y.P.), the National Research Council of Thailand grant (N34A640419), the Centre of Excellence in Clinical Virology of Chulalongkorn University/King Chulalongkorn Memorial Hospital (GCE 59 009-30-005) and the Second Century Fund (C2F), Chulalongkorn University (to J.M.). This research project is supported by grants for development of new faculty Staff, Ratchadaphiseksomphot Fund, Chulalongkorn University; Thailand Science Research and Innovation Fund Chulalongkorn University (CU_FRB65_hea (27)_034_30_15).

Supplementary Materials

Supplementary data can be found with this article online at http://www.genominfo.org.

References

1. Sendfeld F, Selga E, Scornik FS, Perez GJ, Mills NL, Brugada R. Experimental models of Brugada syndrome. Int J Mol Sci 2019;20:2123.
2. Makarawate P, Glinge C, Khongphatthanayothin A, Walsh R, Mauleekoonphairoj J, Amnueypol M, et al. Common and rare susceptibility genetic variants predisposing to Brugada syndrome in Thailand. Heart Rhythm 2020;17:2145-2153.
3. Makarawate P, Chaosuwannakit N, Ruamcharoen Y, Panthongviriyakul A, Pongchaiyakul C, Tharaksa P, et al. Prevalence and associated factors of early repolarization pattern in healthy young northeastern Thai men: a correlation study with Brugada electrocardiography. J Arrhythm 2015;31:215-220.
4. Vutthikraivit W, Rattanawong P, Putthapiban P, Sukhumthammarat W, Vathesatogkit P, Ngarmukos T, et al. Worldwide prevalence of Brugada syndrome: a systematic review and meta-analysis. Acta Cardiol Sin 2018;34:267-277.
5. Barc J, Tadros R, Glinge C, Chiang DY, Jouni M, Simonet F, et al. Genome-wide association analyses identify new Brugada syndrome risk loci and highlight a new mechanism of sodium channel regulation in disease susceptibility. Nat Genet 2022;54:232-239.
6. Le Scouarnec S, Karakachoff M, Gourraud JB, Lindenbaum P, Bonnaud S, Portero V, et al. Testing the burden of rare variation in arrhythmia-susceptibility genes provides new insights into molecular diagnosis for Brugada syndrome. Hum Mol Genet 2015;24:2757-2763.
7. Wilde AA, Behr ER. Genetic testing for inherited cardiac disease. Nat Rev Cardiol 2013;10:571-583.
8. Lee SB, Wheeler MM, Thummel KE, Nickerson DA. Calling star alleles with Stargazer in 28 pharmacogenes with whole genome sequences. Clin Pharmacol Ther 2019;106:1328-1337.
9. Mauleekoonphairoj J, Chamnanphon M, Khongphatthanayothin A, Sutjaporn B, Wandee P, Poovorawan Y, et al. Phenotype prediction and characterization of 25 pharmacogenes in Thais from whole genome sequencing for clinical implementation. Sci Rep 2020;10:18969.
10. Telenti A, Pierce LC, Biggs WH, di Iulio J, Wong EH, Fabani MM, et al. Deep sequencing of 10,000 human genomes. Proc Natl Acad Sci U S A 2016;113:11901-11906.

11. Handley SA. The virome: a missing component of biological interaction networks in health and disease. Genome Med 2016;8:32.
12. Moustafa A, Xie C, Kirkness E, Biggs W, Wong E, Turpaz Y, et al. The blood DNA virome in 8,000 humans. PLoS Pathog 2017;13:e1006292.
13. Grandi N, Tramontano E. Human endogenous retroviruses are ancient acquired elements still shaping innate immune responses. Front Immunol 2018;9:2039.
14. Xue B, Sechi LA, Kelvin DJ. Human endogenous retrovirus K (HML-2) in health and disease. Front Microbiol 2020;11:1690.
15. Kremer D, Gruchot J, Weyers V, Oldemeier L, Gottle P, Healy L, et al. pHERV-W envelope protein fuels microglial cell-dependent damage of myelinated axons in multiple sclerosis. Proc Natl Acad Sci U S A 2019;116:15216-15225.
16. Levet S, Charvet B, Bertin A, Deschaumes A, Perron H, Hober D. Human endogenous retroviruses and type 1 diabetes. Curr Diab Rep 2019;19:141.
17. Song Y, Choi JE, Kwon YJ, Chang HJ, Kim JO, Park DH, et al. Identification of susceptibility loci for cardiovascular disease in adults with hypertension, diabetes, and dyslipidemia. J Transl Med 2021;19:85.
18. Ariza ME, Williams MV. A human endogenous retrovirus K dUTPase triggers a TH1, TH17 cytokine response: does it have a role in psoriasis? J Invest Dermatol 2011;131:2419-2427.
19. Reynier F, Verjat T, Turrel F, Imbert PE, Marotte H, Mougin B, et al. Increase in human endogenous retrovirus HERV-K (HML-2) viral load in active rheumatoid arthritis. Scand J Immunol 2009;70:295-299.
20. Gao Y, Yu XF, Chen T. Human endogenous retroviruses in cancer: expression, regulation and function. Oncol Lett 2021;21:121.
21. Fernandez AF, Esteller M. Viral epigenomes in human tumorigenesis. Oncogene 2010;29:1405-1420.
22. Farrell PJ. Epstein-Barr virus and cancer. Annu Rev Pathol 2019;14:29-53.
23. Stephens Z, O'Brien D, Dehankar M, Roberts LR, Iyer RK, Kocher JP. Exogene: a performant workflow for detecting viral integrations from paired-end next-generation sequencing data. PLoS One 2021;16:e0250915.
24. Thiene G. Sudden cardiac death in the young: a genetic destiny? Clin Med (Lond) 2018;18:s17-s23.
25. Zack F, Klingel K, Kandolf R, Wegener R. Sudden cardiac death in a 5-year-old girl associated with parvovirus B19 infection. Forensic Sci Int 2005;155:13-17.
26. Juhasz Z, Tiszlavicz L, Kele B, Terhes G, Deak J, Rudas L, et al. Sudden cardiac death from parvovirus B19 myocarditis in a young man with Brugada syndrome. J Forensic Leg Med 2014;25:8-13.
27. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 2014;30:2114-2120.
28. 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.
29. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012;9:357-359.
30. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 2012;19:455-477.
31. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics 2009;10:421.
32. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods 2015;12:59-60.
33. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010;26:841-842.
34. Frankish A, Diekhans M, Jungreis I, Lagarde J, Loveland JE, Mudge JM, et al. Gencode 2021. Nucleic Acids Res 2021;49:D916-D923.
35. Jurka J. Repbase update: a database and an electronic journal of repetitive elements. Trends Genet 2000;16:418-420.
36. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. Nature 2012;489:57-74.
37. Nurk S, Bankevich A, Antipov D, Gurevich A, Korobeynikov A, Lapidus A, et al. Assembling genomes and mini-metagenomes from highly chimeric reads. In: Research in Computational Molecular Biology. RECOMB 2013. Lecture Notes in Computed Science, Vol. 7821 (Deng M, Jiang R, Sun F, Zhang X, eds.). Berlin: Springer, 2013. pp. 158-170.
38. Schwarz E, Freese UK, Gissmann L, Mayer W, Roggenbuck B, Stremlau A, et al. Structure and transcription of human papillomavirus sequences in cervical carcinoma cells. Nature 1985;314:111-114.
39. Lee M, Hills M, Conomos D, Stutz MD, Dagg RA, Lau LM, et al. Telomere extension by telomerase and ALT generates variant repeats by mechanistically distinct processes. Nucleic Acids Res 2014;42:1733-1746.
40. Sung WK, Zheng H, Li S, Chen R, Liu X, Li Y, et al. Genome-wide survey of recurrent HBV integration in hepatocellular carcinoma. Nat Genet 2012;44:765-769.
41. Hudnall SD, Chen T, Allison P, Tyring SK, Heath A. Herpesvirus

prevalence and viral load in healthy blood donors by quantitative real-time polymerase chain reaction. Transfusion 2008;48:1180-1187.

42. Zapatka M, Borozan I, Brewer DS, Iskar M, Grundhoff A, Alawi M, et al. The landscape of viral associations in human cancers. Nat Genet 2020;52:320-330.
43. Jurasz H, Pawlowski T, Perlejewski K. Contamination issue in viral metagenomics: problems, solutions, and clinical perspectives. Front Microbiol 2021;12:745076.
44. Xie Y, Xue Q, Jiao W, Wu J, Yu Y, Zhao L, et al. Associations between sputum torque teno virus load and lung function and disease severity in patients with chronic obstructive pulmonary disease. Front Med (Lausanne) 2021;8:618757.
45. Spandole S, Cimponeriu D, Berca LM, Mihaescu G. Human anelloviruses: an update of molecular, epidemiological and clinical aspects. Arch Virol 2015;160:893-908.
46. Desingu PA, Nagarajan K, Dhama K. Can a torque teno virus (TTV) be a naked DNA particle without a virion structure? Fron Virol 2022;2:821298.
47. Sawata T, Bando M, Nakayama M, Mato N, Yamasawa H, Takahashi M, et al. Clinical significance of changes in Torque teno virus DNA titer after chemotherapy in patients with primary lung cancer. Respir Investig 2018;56:173-178.
48. Matsubara H, Michitaka K, Horiike N, Yano M, Akbar SM, Torisu M, et al. Existence of TT virus DNA in extracellular body fluids from normal healthy Japanese subjects. Intervirology 2000;43:16-19.
49. Ninomiya M, Takahashi M, Shimosegawa T, Okamoto H. Analysis of the entire genomes of fifteen torque teno midi virus variants classifiable into a third group of genus Anellovirus. Arch Virol 2007;152:1961-1975.
50. Chen X, Li D. Sequencing facility and DNA source associated patterns of virus-mappable reads in whole-genome sequencing data. Genomics 2021;113:1189-1198.
51. Cantalupo PG, Katz JP, Pipas JM. HeLa nucleic acid contamination in the cancer genome atlas leads to the misidentification of human papillomavirus 18. J Virol 2015;89:4051-4057.
52. Harputluoglu M, Carr BI. Hepatitis B before and after hepatocellular carcinoma. J Gastrointest Cancer 2021;52:1206-1210.
53. Chen X, Kost J, Sulovari A, Wong N, Liang WS, Cao J, et al. A virome-wide clonal integration analysis platform for discovering cancer viral etiology. Genome Res 2019;29:819-830.
54. Wallace AD, Wendt GA, Barcellos LF, de Smith AJ, Walsh KM, Metayer C, et al. To ERV is human: a phenotype-wide scan linking polymorphic human endogenous retrovirus-K insertions to complex phenotypes. Front Genet 2018;9:298.
55. Jacques PE, Jeyakani J, Bourque G. The majority of primate-specific regulatory sequences are derived from transposable elements. PLoS Genet 2013;9:e1003504.
56. Garazha A, Ivanova A, Suntsova M, Malakhova G, Roumiantsev S, Zhavoronkov A, et al. New bioinformatic tool for quick identification of functionally relevant endogenous retroviral inserts in human genome. Cell Cycle 2015;14:1476-1484.
57. Suntsova M, Garazha A, Ivanova A, Kaminsky D, Zhavoronkov A, Buzdin A. Molecular functions of human endogenous retroviruses in health and disease. Cell Mol Life Sci 2015;72:3653-3675.
58. Benito B, Sarkozy A, Mont L, Henkens S, Berruezo A, Tamborero D, et al. Gender differences in clinical manifestations of Brugada syndrome. J Am Coll Cardiol 2008;52:1567-1573.
59. Nielsen MW, Holst AG, Olesen SP, Olesen MS. The genetic component of Brugada syndrome. Front Physiol 2013;4:179.
60. Ohno S, Zankov DP, Ding WG, Itoh H, Makiyama T, Doi T, et al. KCNE5 (KCNE1L) variants are novel modulators of Brugada syndrome and idiopathic ventricular fibrillation. Circ Arrhythm Electrophysiol 2011;4:352-361.
61. Benito B, Berruezo A. Brugada syndrome and pregnancy: delving into the role of sex hormones in ion channelopathies. Rev Esp Cardiol (Engl Ed) 2014;67:165-167.
62. Yang G, Liu J, Wang Y, Du Y, Ma A, Wang T. Lack of influence of sex hormones on Brugada syndrome-associated mutant Nav1.5 sodium channel. J Electrocardiol 2019;52:82-87.
63. Katoh I, Kurata S. Association of endogenous retroviruses and long terminal repeats with human disorders. Front Oncol 2013;3:234.
64. Krasovec T, Sikonja J, Zerjav Tansek M, Debeljak M, Ilovar S, Trebusak Podkrajsek K, et al. Long-term follow-up of three family members with a novel NNT pathogenic variant causing primary adrenal insufficiency. Genes (Basel) 2022;13:717.
65. Williams JL, Hall CL, Meimaridou E, Metherell LA. Loss of Nnt increases expression of oxidative phosphorylation complexes in C57BL/6J hearts. Int J Mol Sci 2021;22:6101.
66. Oakley RH, Cidlowski JA. Glucocorticoid signaling in the heart: a cardiomyocyte perspective. J Steroid Biochem Mol Biol 2015;153:27-34.
67. Nickel AG, von Hardenberg A, Hohl M, Loffler JR, Kohlhaas M, Becker J, et al. Reversal of mitochondrial transhydrogenase causes oxidative stress in heart failure. Cell Metab 2015;22:472-484.
68. Zhu L, Su X. Case report: neuroblastoma breakpoint family genes associate with 1q21 copy number variation disorders. Front Genet 2021;12:728816.
69. Vandepoele K, Andries V, Van Roy N, Staes K, Vandesompele J, Laureys G, et al. A constitutional translocation t(1;17)

(p36.2;q11.2) in a neuroblastoma patient disrupts the human NBPF1 and ACCN1 genes. PLoS One 2008;3:e2207.

70. Andries V, Vandepoele K, Van Roy F. The NBPF gene family. In: Neuroblastoma: Present and Future (Shimada H, ed.). Rijeka: InTech, 2012. pp. 185-214.

71. Kumarswamy R, Bauters C, Volkmann I, Maury F, Fetisch J, Holzmann A, et al. Circulating long noncoding RNA, LIPCAR, predicts survival in patients with heart failure. Circ Res 2014;114:1569-1575.

72. Liao J, He Q, Li M, Chen Y, Liu Y, Wang J. LncRNA MIAT: myocardial infarction associated and more. Gene 2016;578:158-161.

73. Zhu L, Li N, Sun L, Zheng D, Shao G. Non-coding RNAs: the key detectors and regulators in cardiovascular disease. Genomics 2021;113:1233-1246.

74. Wang Y, Hu Y, Wu G, Yang Y, Tang Y, Zhang W, et al. Long noncoding RNA PCAT-14 induces proliferation and invasion by hepatocellular carcinoma cells by inducing methylation of miR-372. Oncotarget 2017;8:34429-34441.

75. Ge S, Mi Y, Zhao X, Hu Q, Guo Y, Zhong F, et al. Characterization and validation of long noncoding RNAs as new candidates in prostate cancer. Cancer Cell Int 2020;20:531.

76. Kullo IJ, Shameer K, Jouni H, Lesnick TG, Pathak J, Chute CG, et al. The ATXN2-SH2B3 locus is associated with peripheral arterial disease: an electronic medical record-based genome-wide association study. Front Genet 2014;5:166.

77. Shukla S, Zhang X, Niknafs YS, Xiao L, Mehra R, Cieslik M, et al. Identification and validation of PCAT14 as prognostic biomarker in prostate cancer. Neoplasia 2016;18:489-499.