Investigation of somatic single nucleotide variations in human endogenous retrovirus elements and their potential association with cancer

Ting-Chia Chang, Santosh Goud, John Torcivia-Rodriguez, Yu Hu, Qing Pan, Robel Kahsay, Jonas Blomberg, Raja Mazumder

1 Department of Biochemistry & Molecular Medicine, George Washington University Medical Center, Washington, DC, United States of America, 2 The School of Systems Biology, George Mason University, Manassas, VA, United States of America, 3 Department of Statistics, The George Washington University, Washington, DC, United States of America, 4 Department of Medical Sciences, Uppsala University Hospital, Uppsala, Sweden, 5 McCormick Genomic and Proteomic Center, George Washington University, Washington, DC, United States of America

These authors contributed equally to this work.

* mazumder@gwu.edu

Abstract

Human endogenous retroviruses (HERVs) have been investigated for potential links with human cancer. However, the distribution of somatic nucleotide variations in HERV elements has not been explored in detail. This study aims to identify HERV elements with an over-representation of somatic mutations (hot spots) in cancer patients. Four HERV elements with mutation hotspots were identified that overlap with exons of four human protein coding genes. These hotspots were identified based on the significant over-representation (p < 8.62e-4) of non-synonymous single-nucleotide variations (nsSNVs). These genes are TNN (HERV-9/LTR12), OR4K15 (HERV-IP10F/LTR10F), ZNF99 (HERV-W/HERV17/LTR17), and KIR2DL1 (MST/MaLR). In an effort to identify mutations that impact survival, all nsSNVs were further evaluated and it was found that kidney cancer patients with mutation C2270G in ZNF99 have a significantly lower survival rate (hazard ratio = 2.6) compared to those without it. Among HERV elements in the human non-protein coding regions, we found 788 HERVs with significantly elevated numbers of somatic single-nucleotide variations (SNVs) (p < 1.60e-5). From this category the top three HERV elements with significantly over-represented SNVs are HERV-H/LTR7, HERV-9/LTR12 and HERV-L/MLT2. Majority of the SNVs in these 788 HERV elements are located in three DNA functional groups: long non-coding RNAs (lncRNAs) (60%), introns (22.2%) and transcriptional factor binding sites (TFBS) (14.8%). This study provides a list of mutational hotspots in HERVs, which could potentially be used as biomarkers and therapeutic targets.
**Background**

Endogenous retroviruses (ERVs) have been embedded in the primate genomes for over 30 million years [1, 2]. Typically, the genetic structure of ERVs contains the internal coding sequencing of the four proviral genes (gag, pro, pol and env) along with two long terminal repeats (LTRs) [3]. Over the course of time, most ERVs in the human genome have been severely damaged in their original genetic structure due to the accumulation of mutations, insertions, deletions and translocations that have spliced out the original coding region of proviral genes between two flanking LTRs [4, 5]. Solitary LTRs are the most common ERVs within the human genome [4, 5].

Human endogenous retroviruses (HERVs) account for approximately eight percent of the human genome [6] and have been classified into three main classes I, II and III. This classification is based on sequence similarity to different genera of infectious retroviruses [7]. Over 22 distinct HERV groups across three classes have been reported [8]. Gamma- and Epsilon-retrovirus like (GE; earlier called Class I) HERVs are linked to gamma-retroviruses like murine leukemia virus (MLV); It includes HERV-W (HERV17/LTR17/ERV-W), HERV-H (HERV-H/LTR7), and other subgroups. The Alpha and Beta-retrovirus like (AB; earlier called Class II) HERVs [9] are related to beta-retroviruses such as mouse mammary tumor virus and include several types of HERV-K (HML families) elements. Spumavirus like (S; earlier called Class III) HERVs are distantly related to spumaviruses and include HERV-L (HERV-L/MLT2) and HERV-S (HERV18/LTR18). Classifications of HERV elements are currently not entirely consistent due to varying approaches used to detect HERV sequences [8, 10]. This leads to a variable number of identified HERV groups based on the bioinformatic methodology and algorithm used and can cause inconsistencies in HERV classification [11, 12]. Recent work performed by Vargiu et al. has been able to systematically identify and classify 3,173 HERVs in the human genome [13], thereby providing some consistency in HERV classification. Of the 3,137 HERVs, 1,214 canonical (homogeneous) and 1,923 non-canonical (heterogenous) HERVs were separately placed into 39 and 31 groups (clades) respectively. This work builds upon a huge volume of previous work on repetitive elements and evolutionary analysis of ERVs [14, 15] [16–18].

Over the course of evolution, majority of HERVs within the human genome have gradually lost their original protein coding functions [19]. However, HERVs have been indirectly linked to the various diseases, including human preimplantation embryogenesis [20], multiple sclerosis[21], cancers [22–24] and neurological disorders [25]. In cancers, for example, Np9, which is encoded by HERV-K (HML groups) elements, has been proposed to be involved in oncogenic mechanisms through the LNX/Numb/Notch pathway [26, 27]. HERV LTRs have been reported to participate in human tumorigenesis by regulating the expression of its adjacent genes [28, 29]. For example, a mutation found in a HERV LTR leads to the activation of syncytin-1 encoded by HERV-W Env with high expression in bladder carcinoma [30]. It has been proposed that somatic mutations are associated with aberrant activation of stem cell-associated retroviruses (SCAR) and with stem cell-like phenotypes of cancer cells, clinical intractability of human malignancies, and increased likelihood of therapy failure and death from cancer [31]

Next-Generation Sequencing (NGS) has immensely aided the identification of genetic variations and their role in human diseases [32–34]. The availability of single nucleotide variation (SNV) databases and locus specific disease-related annotation databases has helped researchers map mutations to potential biomarkers [35]. In this study, we have analyzed such datasets to explore pan-cancer mutations in HERV elements. Although, it has been shown in several studies that such pan-cancer analysis can help identify patterns of driver mutations [36–38], to the best of our knowledge no such study has been performed on HERV elements.
This study explores correlations between HERV elements and cancers by identifying somatic mutation hotspots in the human genome, followed by a detailed review of functional annotations available for these genomic regions.

Material and methods

Data integration

SNV data. All SNV data was retrieved from BioMuta. BioMuta v3.0 [35, 39] is a comprehensive non-redundant data set of SNVs found in cancer within the coding region of hg19/GRCh37 compiled from multiple sources, including The Cancer Genome Atlas (TCGA) [40], CGHub data portal (https://cghub.ucsc.edu/), ClinVar [41], Catalog Somatic Mutations in Cancer (COSMIC) [42], International Cancer Genome Consortium (ICGC) [43], Integrative Onco Genomics (IntOGen) [44], UniProtKB [45], literature mining, and manual curation. For this study, BioMuta v3.0 was expanded to include SNVs in the non-coding regions of the human genome (hg19/GRCh37). The non-coding SNVs mutation data was extracted from ICGC (Version 23) [43] and COSMIC (version 79) [42]. This non-coding database was derived from whole genome sequencing (WGS) data and was restricted to regions excluding coding domain sequences based on the annotations available through UCSC genome browser (https://genome.ucsc.edu/). It is important to note that single nucleotide polymorphisms (SNPs) were filtered out for both coding and non-coding SNV datasets by Mutect 2 for TCGA [46] and by MuTect and Strelka [47] for ICGC [48] and DNA-Seq analysis pipelines for TCGA is listed in (https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/). SNV identification in repetitive regions is handled based on published methods [49–51].

HERV data. Comprehensive HERV data set based on hg19/GRCh37 (S1 Table) was obtained using methods described by Vargiu et al. [13]. The detection of HERV elements is based on three basic principles: (a) detection of candidate LTRs, (b) detection of chains of conserved retroviral motifs fulfilling distance constraints, and (c) an attempt at reconstruction of original retroviral protein sequences, combining alignment, codon statistics and properties of protein ends. All classifications and nomenclature of HERV elements that is used for this study is available in S2 Table. The term canonical HERV is defined to be a HERV sequence that comes from one HERV group [13]. Otherwise, a HERV element is considered non-canonical if composed of two or more HERV groups. The names of groups are based on prior usage in the literature and from RepeatMasker [16].

DNA functional elements data

The DNA functional elements data set for protein non-coding regions is comprised of: tRNA (version 1.23) [52–54]; CpG island (download date: Oct. 2015) (UCSC Genome Browser); Open Regulatory Annotations (version 3.0) [55, 56]; microRNA sites and their target sites (version 7.1) [57]; pseudo exons (release 60) [58, 59]; VISTA enhancers (download date: Oct. 2015) [60], HACA box elements; CDBox elements; SnoRNA/ScaRNA sites (version 3) [61], alternative splice sites (download date: Oct. 2015) (UCSC genome Browser); transcription factor binding sites (version 3.0) [62, 63], introns (download date: Oct. 2015) (UCSC genome Browser), and long non-coding RNAs (v4.0) [64].

Over-representaion analysis

SNVs were mapped to HERV elements based on their genomic coordinates. SNVs could be either from human protein coding region or protein non-coding region. Retrovirus coding region such as gag, pro, pol and env are not considered as human protein coding region unless...
there is annotation in UniProtKB/Swiss-Prot [45] that indicates that they are indeed part of a protein coding sequence. The Binomial test [65, 66] was then used to evaluate the significance of the over-represented SNVs in each HERV element, by comparing its observed SNV number to the expected SNV number in each HERV element on human protein coding region or non-coding region, respectively [67, 68]. The calculation of the expected number n(E) of SNV sites in each HERV element is expressed as follows:

\[ p(F) = \frac{n(F)}{L} \]

Where \( n(F) \) is the total number of nucleotides at each HERV element (total base pairs of each HERV element) and \( L \) is the total number of nucleotides of the genome (total base pairs of human chromosomal protein coding or non-coding region). The probability \( p(F) \) of observing a nucleotide from the human genome at a certain HERV element is calculated by taking the value of \( n(F) \) divided by total length of human chromosomal protein coding or protein non-coding genome \( L \).

\[ n(E) = N \times p(F) = N \times \frac{n(F)}{L} \]

Here, \( N \) is the total number of variation sites found in the human protein coding or non-coding part of the genome. Assuming that somatic SNV sites are equally likely to be found along the entire genome, the value of \( n(E) \) gives us the expected number of SNV sites that would be found in the HERV regions of the genome. The expected ratio in whole human genome is 0.019 (total number of SNVs in coding and non-coding region divided by total number of bases in the human genome). To evaluate the expected ratio in the whole genome, random sampling of permutation [69] was performed in R (http://www.R-project.org/) for comparing the observed ratio in random fragments and calculating the number of SNVs in each fragment (1000 bases in one fragment). (S3 Table)

Binomial statistic was used to calculate the p-value of the expected versus observed as follows:

\[ P-value = \sum_{n=0}^{N} \left( \frac{N!}{n!(N-n)!} \right) \times p(F)^n \times (1-p(F))^{N-n} \]

Where \( n(O) \) is the observed number of SNV sites within the HERV elements being examined. Bonferroni correction was used to calculate the threshold for the p-Value (0.05/n) (n represents the number of Binomial tests we performed here). The p-value used as our significance cutoff was 8.62e-4 (0.05/58) for the protein coding region nsSNVs (S4 Table). For protein non-coding region SNVs in HERVs, the significance cutoff was 1.60e-5 (0.05/3,130) (S5 Table).

**Differential expression analysis**

BioXpress is a gene expression database, which provides differential expression of both gene and miRNA in cancer [39, 70]. With respect to a specific cancer type, differential expression analysis using DESeq2 [71] is performed on expression levels of each gene or miRNA in tumor and adjacent non-tumor samples. Current version of BioXpress includes 34 TCGA cancer types (mapped to 73 DOIDs (Cancer Disease Ontology IDs)) [72], in which 20,502 genes and 1,965 miRNAs were analyzed and 18,846 genes and 710 miRNAs have been observed to be significantly differentially expressed in at least one cancer type. Genes of interest (HERVs that overlap with protein coding genes and are somatic mutation hotspots), identified in this study, were further explored in BioXpress to find out if they are significantly overexpressed in cancer.
Survival analysis

Key nsSNVs that were identified were further investigated by log-rank test to evaluate their possible impact on patient survival. Patient clinical information was retrieved from TCGA (https://portal.gdc.cancer.gov/). Each key nsSNV was retrieved which have significant (p < 0.05) differential expression in the certain cancer types. BioMuta nsSNVs from non-TCGA sources was removed since BioXpress only uses TCGA data. Cancer patients were divided into two groups: one group where the patients have the key nsSNV and the other where they do not. Log-rank test was applied to test the death time distributions for the two groups. The Cox model was used to adjust for factors such as age at initial diagnosis, pathological stage and gender. SAS (version 9.3) using previously published method was used to perform the analysis [73].

Results

Genome-wide identification of somatic SNVs in HERV elements

In the protein coding region of the human genome 2,867,887 sites impacted by somatic nsSNVs were found. Whereas, in human protein non-coding region, 59,205,289 SNV sites with somatic SNVs were identified (S6 and S7 Tables). To confirm the coverage of SNVs in whole human genome, all SNVs were mapped to the genome. S1 File shows the distributions of all SNVs in chromosome 1 to 22, X, Y. The gap in the plot represents the repeat sequence and low complexity centromeric region where SNVs are hard to identify.

A total of 2,543 somatic nsSNVs were identified in 25 HERV groups that overlap with human protein coding regions (S1 and S6 Tables). Amongst them, 919 nsSNVs were identified in seven Gamma-retrovirus/Epsilon-retrovirus-related (GE) canonical HERV groups. The rest 1,624 nsSNVs were found in 20 non-canonical HERV groups. Ten of the groups were from Gamma-retrovirus/Epsilon-retrovirus-related (GE) non-canonical retrovirus, which involved 50.1% of the total examined nsSNVs. Five out of 20 groups belonging to Alpha-retrovirus/Beta-retrovirus-related (AB) retroviruses contained 23.5% of the nsSNVs. 20.4% of nsSNVs were within two Spumavirus related (S) HERV groups. 5.8% of the nsSNVs were found in the "Uncertain Errantivirus-like" group and two unclassified HERV groups.

To investigate the distribution of somatic SNVs in HERV elements in protein non-coding regions, we mapped SNVs present in protein non-coding genomic regions HERV genomic coordinates. The results indicate 433,409 human chromosomal non-coding SNVs are located in HERV elements. 167,561 of them (38.7%) were in canonical HERV groups (Table 1 and S7 Table). Of these 167,561 mutations, 135,032, or 80.8% were found within HERV Gamma-retrovirus/Epsilon-retrovirus-related (GE) retroviruses. 18,295 (10.9%), and 14,243 (8.3%) were found in HERV Alpha-retrovirus/Beta-retrovirus-related (AB) retroviruses, and Spumavirus-related (S) retroviruses, respectively. The rest 265,848 SNVs were found in the non-canonical HERV groups. The non-canonical group classifications are Gamma-retrovirus/Epsilon-retrovirus-related (GE) retroviruses, Alpha-retrovirus/Beta-retrovirus-related (AB) retroviruses, Spumavirus-related (S) retroviruses, Uncertain Errantivirus-like proviruses, and unclassified groups. The proportion of mutations found in these groups are 72.6% (193,268/265,848), 18.5% (49,324/265,848), 8.3% (22,043/265,848), < 0.001% (89/265,848), and 0.4% (1,124/265,848) respectively (S7 Table).

Somatic SNV hotspots in HERV elements

To understand whether certain HERV elements contain a significantly different number of SNVs than expected, we compared the observed mutations within each HERV element to expected number of mutations.
SNVs from human protein coding region with HERV elements. There are several HERV elements that overlap with the exon of genes. We identified HERVs which contained significantly more nsSNVs than expected ($p < 8.62e^{-4}$). There were 83 HERVs with 2,543 nsSNVs (Fig 1A). 39% of the elements (32/83), encompassing 1,855 of the 2,543 nsSNVs, have significant numbers of nsSNVs ($S4$ Table). Majority of the HERV elements (28/32, 87.5%) are in 18 different HERV groups, and of these, 28 HERVs have significantly fewer nsSNVs than expected in the protein coding region. The remaining 4 (12.5%) HERV elements include significant over-representation of nsSNVs ($Table 2$ and $S8$ Table) containing 492 of the 1,855 (26%) and hence was prioritized for further functional analysis. In total, 492 nsSNVs are concentrated in the exon regions of 4 genes: tenasin N (TNN) (with HERV-9/LTR12) ($p = 6.67e^{-4}$), zinc finger protein 99 (ZNF99) (with HERV-W/HERV17/LTR17) ($2.63e^{-26}$), killer cell immunoglobulin like receptor, two Ig domains and long cytoplasmic tail 1 (KIR2DL1) (with MaLR/MST) ($3.26e^{-7}$), and olfactory receptor family 4 subfamily K member 15 (OR4K15) (with HERV-IP10F/LTR10F) ($2.58e^{-5}$) ($Table 2$, $S8$ and $S9$ Tables).

SNVs from human protein non-coding regions mapping to HERV elements. There are 1,820 HERV elements in protein non-coding regions which contain significant over- or under-representation of SNVs (Fig 1B and $S5$ Table) ($p < 1.60e^{-5}$). Most of the HERV elements (1,032 or 57%) have significantly fewer mutations than expected. The remaining 788 HERV elements (43%) have an over-representation of SNVs, which account for more than 73% of the total SNVs. 193,439 SNVs ($S5$ and $S10$ Tables) were identified in 788 HERV elements that contained a significantly high number of mutations ($p < 1.60e^{-5}$). The proportion of somatic mutations in canonical HERVs (41.7%) is lower than that of non-canonical HERVs (58.3%). For canonical HERVs, the proportion of somatic mutations in Gamma-retrovirus/Epsilon-retrovirus-related (GE) retroviruses, Alpha-retrovirus/Beta-retrovirus-related (AB) retroviruses, Spumavirus-related (S) retroviruses HERV groups are 80.9%, 9.3%, 9.8%, respectively. For non-canonical HERVs, the proportion of mutations in classes Gamma-retrovirus/Epsilon-retrovirus-related (GE) retroviruses, Alpha-retrovirus/Beta-retrovirus-related (AB) retroviruses, Spumavirus-related (S) retroviruses are 72.9%, 17.3%, and 9.8% respectively.

Genome-wide pan-cancer analysis

HERV elements with over-representation of nsSNVs. To further investigate the impact of nsSNVs in HERV regions in different cancer types, cancer terms were unified using cancer
The relationship between nsSNVs in the protein coding region of HERV elements and multiple cancer types is shown in Fig 2. It can be seen from Table 2 that 81.3% of the nsSNVs are from Gamma-retrovirus/Epsilon-retrovirus-related (GE) non-canonical HERV elements (HERV-W/LTR17/HERV17, HERV-9/LTR12 and HERV-IP10F/LTR10F). These nsSNVs are associated with at least 20 cancer types. The top three cancer types—skin cancer, lung cancer, and colon cancer—are associated with mutated sites in HERV-W/LTR17/HERV17. The proportion of mutations in HERV-W/LTR17/HERV17 was 50.6% (249/492). HERV element in Spumavirus-related (S) (MaLR/MST) is found in several cancer types. The proportion of mutations in (MaLR/MST) was 18.7% (92/492). This set of MaLR/MST mutations is found in many cancer types including skin cancer, thyroid cancer, and lung cancer.

HERV elements with an over-representation of SNVs from protein non-coding regions. SNVs in the human chromosomal non-coding region of HERV elements could also lead to carcinogenesis if they impact regulatory regions or protein binding sites [74]. Liang et al. suggested the HERV elements with unstable genomic variants near lncRNA can trigger onset of human adenocarcinoma [75]. Based on our results (Fig 1B), we identified 788 HERV elements with significant over-representation of nsSNVs. S1 Fig indicates each of 788 HERV elements contains 144 SNVs on average, which are coming from skin cancer patient samples. The second highest number (29 SNVs on average) is from esophageal cancer. And the third highest number is from liver cancer (15 SNVs on average).

Fig 3 shows the distribution of SNVs in 357 canonical and 431 non-canonical HERV elements that map to protein non-coding regions. In both canonical (Fig 3A) and non-canonical (Fig 3B) Gamma-retrovirus/Epsilon-retrovirus-related (GE) HERVs, HERV-H/LTR7 and HERV-9/LTR12 contain the largest numbers of SNVs from skin cancer, esophageal cancer, and breast cancer. In canonical Alpha-retrovirus/Beta-retrovirus-related (AB) HERVs, HML-8/HERV-K11I/MER11A contain the most somatic mutations. HML-3/HERV-K9I/MER9 is the most affected non-canonical Alpha-retrovirus/Beta-retrovirus-related (AB) HERVs. In both canonical and non-canonical Spumavirus-related (S) HERVs, HERV-L/LTR7 has the highest SNV sites.

Table 2. Four HERV elements with significantly over-represented nsSNVs mapped to four genes.

<table>
<thead>
<tr>
<th>HERV Sequence</th>
<th>HERV ID</th>
<th>HERV groups</th>
<th>Gene</th>
<th>Chromosome</th>
<th>Length</th>
<th>Mutations</th>
<th>Expected mutations</th>
<th>Difference</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-canon</td>
<td>6114</td>
<td>HERV-9/LTR12</td>
<td>TNN</td>
<td>1</td>
<td>264</td>
<td>35</td>
<td>20.47</td>
<td>14.52</td>
<td>6.67e-04</td>
</tr>
<tr>
<td>Non-canon</td>
<td>4780</td>
<td>MST/MaLR</td>
<td>KIR2DL1</td>
<td>19</td>
<td>702</td>
<td>92</td>
<td>54.43</td>
<td>37.56</td>
<td>3.26e-07</td>
</tr>
<tr>
<td>Non-canon</td>
<td>4062</td>
<td>HERV-IP10F/LTR10F</td>
<td>OR4K15</td>
<td>14</td>
<td>1044</td>
<td>116</td>
<td>80.95</td>
<td>35.04</td>
<td>2.58e-05</td>
</tr>
<tr>
<td>Non-canon</td>
<td>4673</td>
<td>HERV-W/LTR17/HERV17</td>
<td>ZNF99</td>
<td>19</td>
<td>1585</td>
<td>249</td>
<td>122.91</td>
<td>126.08</td>
<td>2.63e-26</td>
</tr>
</tbody>
</table>

https://doi.org/10.1371/journal.pone.0213770.g001
https://doi.org/10.1371/journal.pone.0213770.t002
Expression of genes containing HERV element with over-represented nsSNVs.

Even though the nsSNVs in our dataset are somatic mutations, the correlation between gene expression, HERVs, and each cancer type can provide insights into the role of these genes in cancer. Our results show that nsSNVs found in multiple cancer types are located within exomes of *TNN*, *KIR2DL1*, *OR4K15*, and *ZNF99* genes. We examined whether or not their differential expression were significant in 34 cancer types.
We found that TNN was significantly over-expressed in kidney renal clear cell carcinoma (KIRC) \( (p = 0.024) \) and prostate cancer \( (p = 0) \), but significantly under-expressed in liver \( (p = 0.003) \) and breast cancer \( (p = 0.003) \). 68% (49/72) of patients with KIRC and 84.6% (44/52) of patients with prostate cancer have the same trend of over-expressed TNN (Fig 4). Conversely, 88% (44/50) of patients with liver cancer and 54.4% (62/114) of patients with breast cancer have a similar tendency of under-expressed TNN. Additionally, KIR2DL1 was significantly over-expressed in kidney cancer \( (p \approx 0) \) while significantly under-expressed in lung cancer \( (p \approx 0) \). 91.7% (66/72) patients had a positive correlation of up-regulated KIR2DL1 in kidney cancer. 94.1% (48/51) patients had a direct correlation of down-regulated KIR2DL1 in lung cancer (Fig 4). Moreover, ZNF99 was significantly under-expressed in kidney cancer \( (p \approx 0) \). 84.72% (61/72) of patients with kidney have the same trend of under-expression of ZNF99 in kidney cancer (Fig 4).

**SNVs impact on HERV region and protein functional sites.** In addition to our gene expression analysis, we counted the numbers of nsSNVs that may have an impact on functional sites within these four genes. S11 Table provides the total number of nsSNVs affecting post-translational modification of amino acids. In the gene TNN, cancer-associated nsSNVs within HERV elements were found to likely impact amino acid post-translational modification...
PTM) sites that led to phosphorylation gain or loss. In KIR2DL1, mutations were found to likely impact multiple functions such as phosphorylation, glycosylation, and ligand binding site. In OR4K15 and ZNF99, nsSNVs affect the modification of amino acid phosphorylation and glycosylation.

We combined the results of nsSNVs found in cancers with significant differential gene expression data and impacted PTM sites (Table 3). Three nsSNVs in TNN are found in kidney cancer and breast cancer samples. One of these three nsSNVs impacts a PTM site of the protein. 12 nsSNVs in KIR2DL1 are found in lung cancer and kidney cancer samples. Six out of these 12 nsSNVs affect PTM sites on this protein. In ZNF99, seven nsSNVs are found in kidney cancer and there are no PTM site affecting nsSNVs.

**Survival analysis on 22 key nsSNV in cancer patients.** The total number of nsSNVs in genes which showed significant differential expression in cancer tissues compared to non-tumor tissues were 22 (Table 3). After extracting metadata and clinical data for patients for these 22 nsSNVs, we found one SNV in ZNF99 to be associated closely with patient survival rate, as shown in Fig 5.

As shown in Fig 5, patients with this nsSNV in ZNF99 at amino acid position 757 (A to G modification) have a lower survival rate than the patients without this variation. This key amino acid located in position 757 of ZNF99, which triggers abnormal significant under-expression of ZNF99 in kidney cancer. Additionally, the kidney cancer patients with this key mutation in ZNF99 have a significant decrease in survival rate (Hazard ratio = 2.642; p = 0.05). We believe this nsSNV could be involved in the progression of cancer and further analysis is warranted to validate this mutation.
<table>
<thead>
<tr>
<th>Gene</th>
<th>Chromosome</th>
<th>Position_ N^a</th>
<th>Ref_ N^b</th>
<th>Var_ N^c</th>
<th>Position_ A^d</th>
<th>Ref_ A^e</th>
<th>Var_ A^f</th>
<th>Predicted impact on function</th>
<th>DOID /Cancer type</th>
<th>HERV Id</th>
<th>HERV group</th>
<th>HERV Supergroup</th>
</tr>
</thead>
<tbody>
<tr>
<td>TNN</td>
<td>1</td>
<td>2061</td>
<td>C</td>
<td>A</td>
<td>687</td>
<td>H</td>
<td>Q</td>
<td>-</td>
<td>DOID:1612 / breast cancer</td>
<td>6114</td>
<td>HERV-9/ LTR12</td>
<td>GE^g</td>
</tr>
<tr>
<td>TNN</td>
<td>1</td>
<td>2062</td>
<td>G</td>
<td>A</td>
<td>688</td>
<td>V</td>
<td>M</td>
<td>-</td>
<td>DOID:1612 / breast cancer</td>
<td>6114</td>
<td>HERV-9/ LTR12</td>
<td>GE</td>
</tr>
<tr>
<td>KIR2DL1</td>
<td>19</td>
<td>64</td>
<td>C</td>
<td>A</td>
<td>22</td>
<td>H</td>
<td>N</td>
<td>Gain</td>
<td>Phosphorylation</td>
<td>DOID:1324 / lung cancer</td>
<td>4780</td>
<td>MST/ MaLR</td>
</tr>
<tr>
<td>KIR2DL1</td>
<td>19</td>
<td>95</td>
<td>T</td>
<td>C</td>
<td>32</td>
<td>L</td>
<td>P</td>
<td>-</td>
<td>DOID:1324 / lung cancer</td>
<td>4780</td>
<td>MST/ MaLR</td>
<td>S</td>
</tr>
<tr>
<td>KIR2DL1</td>
<td>19</td>
<td>328</td>
<td>C</td>
<td>T</td>
<td>110</td>
<td>Q</td>
<td>X</td>
<td>-</td>
<td>DOID:1324 / lung cancer</td>
<td>4780</td>
<td>MST/ MaLR</td>
<td>S</td>
</tr>
<tr>
<td>KIR2DL1</td>
<td>19</td>
<td>55</td>
<td>G</td>
<td>T</td>
<td>19</td>
<td>A</td>
<td>S</td>
<td>-</td>
<td>DOID:1324 / lung cancer</td>
<td>4780</td>
<td>MST/ MaLR</td>
<td>S</td>
</tr>
<tr>
<td>KIR2DL1</td>
<td>19</td>
<td>487</td>
<td>G</td>
<td>T</td>
<td>163</td>
<td>E</td>
<td>X</td>
<td>-</td>
<td>DOID:1324 / lung cancer</td>
<td>4780</td>
<td>MST/ MaLR</td>
<td>S</td>
</tr>
<tr>
<td>KIR2DL1</td>
<td>19</td>
<td>506</td>
<td>G</td>
<td>A</td>
<td>169</td>
<td>R</td>
<td>H</td>
<td>-</td>
<td>DOID:1324 / lung cancer</td>
<td>4780</td>
<td>MST/ MaLR</td>
<td>S</td>
</tr>
<tr>
<td>KIR2DL1</td>
<td>19</td>
<td>308</td>
<td>C</td>
<td>A</td>
<td>103</td>
<td>S</td>
<td>Y</td>
<td>Gain</td>
<td>Phosphorylation</td>
<td>DOID:1324 / lung cancer</td>
<td>4780</td>
<td>MST/ MaLR</td>
</tr>
<tr>
<td>KIR2DL1</td>
<td>19</td>
<td>310</td>
<td>G</td>
<td>C</td>
<td>104</td>
<td>V</td>
<td>L</td>
<td>Gain</td>
<td>Phosphorylation</td>
<td>DOID:263 / kidney cancer</td>
<td>4780</td>
<td>MST/ MaLR</td>
</tr>
<tr>
<td>KIR2DL1</td>
<td>19</td>
<td>487</td>
<td>G</td>
<td>A</td>
<td>163</td>
<td>E</td>
<td>K</td>
<td>-</td>
<td>DOID:263 / kidney cancer</td>
<td>4780</td>
<td>MST/ MaLR</td>
<td>S</td>
</tr>
<tr>
<td>KIR2DL1</td>
<td>19</td>
<td>608</td>
<td>A</td>
<td>G</td>
<td>203</td>
<td>H</td>
<td>R</td>
<td>Gain</td>
<td>Phosphorylation</td>
<td>DOID:1324 / lung cancer</td>
<td>4780</td>
<td>MST/ MaLR</td>
</tr>
<tr>
<td>KIR2DL1</td>
<td>19</td>
<td>670</td>
<td>C</td>
<td>A</td>
<td>224</td>
<td>P</td>
<td>T</td>
<td>Gain</td>
<td>Glycosylation</td>
<td>DOID:1324 / lung cancer</td>
<td>4780</td>
<td>MST/ MaLR</td>
</tr>
<tr>
<td>KIR2DL1</td>
<td>19</td>
<td>680</td>
<td>G</td>
<td>T</td>
<td>227</td>
<td>S</td>
<td>I</td>
<td>Gain</td>
<td>Glycosylation</td>
<td>DOID:1324 / lung cancer</td>
<td>4780</td>
<td>MST/ MaLR</td>
</tr>
<tr>
<td>ZNF99</td>
<td>19</td>
<td>1858</td>
<td>C</td>
<td>G</td>
<td>620</td>
<td>P</td>
<td>A</td>
<td>-</td>
<td>DOID:263 / kidney cancer</td>
<td>4673</td>
<td>HERV-W/ LTR17/ HERV17</td>
<td>GE</td>
</tr>
<tr>
<td>ZNF99</td>
<td>19</td>
<td>2272</td>
<td>G</td>
<td>C</td>
<td>758</td>
<td>E</td>
<td>Q</td>
<td>-</td>
<td>DOID:263 / kidney cancer</td>
<td>4673</td>
<td>HERV-W/ LTR17/ HERV17</td>
<td>GE</td>
</tr>
<tr>
<td>ZNF99</td>
<td>19</td>
<td>2338</td>
<td>A</td>
<td>G</td>
<td>780</td>
<td>K</td>
<td>E</td>
<td>-</td>
<td>DOID:263 / kidney cancer</td>
<td>4673</td>
<td>HERV-W/ LTR17/ HERV17</td>
<td>GE</td>
</tr>
<tr>
<td>ZNF99</td>
<td>19</td>
<td>1557</td>
<td>G</td>
<td>C</td>
<td>519</td>
<td>K</td>
<td>N</td>
<td>-</td>
<td>DOID:263 / kidney cancer</td>
<td>4673</td>
<td>HERV-W/ LTR17/ HERV17</td>
<td>GE</td>
</tr>
<tr>
<td>ZNF99</td>
<td>19</td>
<td>1630</td>
<td>A</td>
<td>T</td>
<td>544</td>
<td>K</td>
<td>X</td>
<td>-</td>
<td>DOID:263 / kidney cancer</td>
<td>4673</td>
<td>HERV-W/ LTR17/ HERV17</td>
<td>GE</td>
</tr>
<tr>
<td>ZNF99</td>
<td>19</td>
<td>2270</td>
<td>C</td>
<td>G</td>
<td>757</td>
<td>A</td>
<td>G</td>
<td>-</td>
<td>DOID:263 / kidney cancer</td>
<td>4673</td>
<td>HERV-W/ LTR17/ HERV17</td>
<td>GE</td>
</tr>
</tbody>
</table>

(Continued)
Distribution of SNVs in HERVs from human protein non-coding regions that overlap with DNA functional elements

To investigate the potential functional roles of the HERV elements from protein non-coding regions, with over-represented mutations, we mapped them to our DNA functional elements dataset. We found that 62,575 variants occur in both HERV elements and in at least one DNA functional element (Fig 6). In our results (Fig 6 and S12 Table), the proportion of SNVs within lncRNA, intron, and TFBS is over 97% of all functionally affected SNVs. The top functional element is long non-coding RNA (lncRNA) which contains 60% of these variants. Additionally, other highly represented functional elements are intron (22.2%), transcription factor

Table 3. (Continued)

<table>
<thead>
<tr>
<th>Gene</th>
<th>Chromosome</th>
<th>Position_Na</th>
<th>Ref_Nb</th>
<th>Var_Nc</th>
<th>Position_Ad</th>
<th>Ref_Ae</th>
<th>Var_Af</th>
<th>Predicted impact on function</th>
<th>DOID /Cancer type</th>
<th>HERV_Id</th>
<th>HERV group</th>
<th>HERV Supergroup</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZNF99</td>
<td>19</td>
<td>1582</td>
<td>A</td>
<td>C</td>
<td>528</td>
<td>K</td>
<td>Q</td>
<td>-</td>
<td>DOID:263 / kidney cancer</td>
<td>4673</td>
<td>HERV-W/LTR17/HERV17</td>
<td>GE</td>
</tr>
</tbody>
</table>

*a Position of nucleotide  
b Reference of nucleotide  
c Variant of nucleotide  
d Position of amino acid  
e Reference of amino acid  
f Variant of nucleotide  
g Gamma-retrovirus/Epsilon-retrovirus-related  
h Spumavirus-related

https://doi.org/10.1371/journal.pone.0213770.t003

Fig 5. The survival rate in the kidney cancer patients with or without the nsSNV at genomic position 2270 (C to G modification) of ZNF99. Kaplan-Meier plot of kidney cancer patient survival based on the existence of ZNF99 (C2270G). X-axis indicate days of survival. Y-axis indicates survival probability. Red and blue lines indicate survival time of kidney cancer patients with and without such mutation respectively. Log-rank test shows that, comparing with patients with ZNF99(C2270G), patients without this nsSNV survive longer with adjusted p-value of 0.05. The hazard ratio is 2.642.

https://doi.org/10.1371/journal.pone.0213770.g005
binding site (TFBS) (15%), alternative splice site (AS) (1.7%), pseudo exon (PE) (0.27%), and CpG Island (0.2%).

Cancer and HERV associations were further explored in lncRNA, intron, and TFBS. Fig 7 shows that, HERV mutations in lncRNAs are found in multiple cancers, especially, skin and esophageal cancer. Additionally, HERV mutation hotspots in introns are from at least 10 cancer types. HERV mutations in TFBS is primarily found in skin cancer and breast cancer;

Fig 6. HERV elements on human non-coding region with over-represented SNVs distributed to DNA functional elements. X-axis is different DNA functional elements. Y-axis is Log2Counts (Counts mean the number of SNVs). The proportion of SNVs on DNA functional elements are LncRNA (long non-coding RNA) (60%), Intron (22.2%), TFBS (transcription factor binding site) (15%), AS (alternative splice site) (1.7%), PE (pseudo exon) (0.27%), and CpG (CpG Island) (0.2%).

https://doi.org/10.1371/journal.pone.0213770.g006

Fig 7. The correlation between HERV-involved DNA functional elements and cancers. Y-axis indicates the SNVs within each HERV class located on each DNA functional elements from top to bottom: SNVs located on HERV elements and DNA functional elements which have combinations containing IncRNA-Can-CI, IncRNA-Can-CII, IncRNA-Can-CIII, IncRNA-Non-CI, IncRNA-Non-CII IncRNA-Non-CIII, Intron-Can-CI, Intron-Can-CII, Intron-Can-CIII, Intron-Non-CI, Intron-Non-CII, Intron-Non-CIII, TFBS-Can-CI, TFBS-Can-CII, TFBS-Can-CIII, TFBS-Non-CI, TFBS-Non-CII, TFBS-Non-CIII, AS-Can-CI, AS-Can-CII, AS-Can-CIII, AS-Non-CI, AS-Non-CII, AS-Non-CIII, PE-Can-CI, PE-Can-CII, PE-Can-CIII, PE-Non-CI, PE-Non-CII, PE-Non-CIII, CpG-Can-CI, CpG-Can-CII, CpG-Can-CIII, CpG-Non-CI, CpG-Non-CII, and CpG-Non-CIII. (Can indicates canonical HERV; Non indicates non-conical HERV; CI indicates Gamma-retrovirus/Epsilon-retrovirus-related HERV; CII indicates Alpha-retrovirus/ Beta-retrovirus-related HERV; CIII indicates Spumavirus-related HERV; LncRNA indicates long non-coding RNA; TFBS indicates transcription factor binding site; AS indicates alternative splice site; PE indicates pseudo exon; CpG indicates CpG Island) HERV-involved lncRNA, intron, and TFBS are found in skin cancer, esophageal cancer, and liver cancer in either canonical HERVs and non-canonical HERVs.

https://doi.org/10.1371/journal.pone.0213770.g007
especially, the non-canonical Gamma-retrovirus/Epsilon-retrovirus-related (GE) HERV in the TFBS. This study provides a direction to narrow down the HERVs in well-defined DNA functional elements which potentially could play a role in cancer.

Discussion

In this study, we provide a large-scale and systematic analysis of somatic SNVs in different HERV classes and their subgroups. In our study, we identified four HERV elements with mutation hotspots that overlap with exons of four genes. Those genes are TNN, KIR2DL1, ZNF99, and OR4K15. It is well known that LTRs have played a significant role in human gene evolution [76]. Yu et al. have suggested that the mutation hotspot located on the 3'-LTR of HERV-W element may have a regulatory role and might be involved in the activation of neighboring genes and its abnormal expression [30]. A few studies support the involvement of TNN and KIR2DL1 in tumorigenesis. TNN is an extracellular matrix protein. It regulates TGF-beta to drive breast carcinogenesis and facilitates the migration of cancer cells into bone [77]. KIR2DL1+ - HLA-C2+ (human leukocyte antigen) genotype was found in oral cancer patients [78]. In Fig 2, multiple exons (Exon 1–5) of KIR2DL1, which overlap with MaLR/MST element seems associated with oral cancer. A study indicated that the ZNF99 gene is involved in the transcription of HERV-W/HERV17/LTR17, which has been implicated in the pathogenesis of multiple sclerosis [79]. Interestingly, the patients with multiple sclerosis have an increased risk of lung, liver, hematologic, and bladder cancer [80]. In Fig 2, ZNF99 corresponding to HERV-W/HERV17/LTR17 are found in few cancers; especially, lung cancer. Importantly, patients with the alteration of amino acid position 757 (A to G modification) in ZNF99 have a lower survival rate than the patients without this variation as shown in Fig 5. Overall, our data indicated over-representative SNVs (hotspots) strengthen the relationship between these four possible HERV-involved genes and multiple cancer types.

We found the SNVs within the HERV-H/LTR7 element are the major family in Gamma-retrovirus/Epsilon-retrovirus-related (GE) retroviruses—both canonical and non-canonical HERVs (Fig 3). Liang et al [75] indicates that non-env-related transcripts of HERV –H are up-regulated in colon cancer cell lines because of abnormal methylation. A few SNVs in HERV-H/LTR7—both canonical and non-canonical—are associated with colon cancer and provides evidence that HERV-H/LTR7 could be involved in tumorigenesis. Additionally, the HML-2/HERV-K/LTR5 element, which is of Alpha-retrovirus/Beta-retrovirus-related (AB) retroviruses, appears in the blood of patients with breast and lymphoma cancer [81]. Based on the results and relevant research, HERV elements on the non-coding region may play a crucial role in cancer.

SNVs in regulatory DNA elements could broadly affect transcription by altering enhancer and promoter activity or chromatin states, leading to abnormal expression in diseases [82]. Further research supports DNA functional elements such as IncRNAs could be involved in carcinogenesis because mutations are located in the DNA functional region [83, 84]. Several SNVs are implicated in the expression of cancer-associated IncRNAs—including CCAT2 in colorectal cancers [85] and PCAT-1 in prostate cancer [86]. Additionally, gamma-like retrovirus, one virus of Gamma-retrovirus/Epsilon-retrovirus-related (GE) HERVs, has a connection to carcinogenesis through IncRNA [87]. HERVs that map to IncRNAs which contain multiple somatic SNVs could play a potential role in carcinogenesis, especially, skin and esophageal cancer. Similarly HERVs mapped to introns have been implicated with cancer formation [88]. This was the first report that an intronic mutation was related to the development of cancer [88]. Moreover, copy number variants of the TFBS which is involved in a proliferation effector-gene and an apoptosis effector-gene are highly associated with melanoma and breast cancer [88].
One of the limitations is the lack of comprehensive SNV datasets for protein non-coding regions. Due to the cost of human whole genome sequencing (WGS) in cancer patients, majority of the data is from whole exome sequencing (WEX) instead of WGS. In order to overcome inherent limitations in the data, more functional analysis is required for supporting the potential mechanism of HERV elements’ association with cancers. Another limitation is the mutation calling within repetitive region. To effectively minimize issues that could potentially cause bias in the sequencing result, the review offers insightful perspectives and potential solutions in the alignment step [49]. Firstly, the pipeline of TCGA and other resources that we used adopt similar mate-pair information from reads that were sequenced in pairs for alignment steps to account for repetitive regions in human genome [50]. Duplicate reads are also marked after alignment, sorting and merging via Picard Tools MarkDuplicate function. This is based on the similarity of the 5’ and 3’ ends of the strands. This is used to protect against a single region being sequenced in much higher quantity than the genome at large; but also help mitigate repetitive regions by essentially labeling them as duplicates and then removing them from consideration. This general same principle also applies to repetitive regions to decrease the chance that all the reads get mapped to one region and variants are called off of that even though the reads are from different very similar regions. Another strategy for handling repeats “is to compute statistics on the depth of coverage for each contig” quoted from Treangen’s paper [49]. The TCGA pipeline checks for higher than 50X sequence coverage for mutation calling (https://docs.gdc.cancer.gov/Data/Introduction/). TCGA and ICGC data has been used to study regulatory elements including untranslated regions, splice sites and non-coding RNA with repeats [51]. We believe that although there is a possibility that the repeats might result in misalignments, the methods used to determine these SNVs are robust enough to provide reliable mutation calls and calculate SNV hotspots.

This study has identified mutational hotspots in HERVs and attempts to rank HERVs which might be associated with cancer. Although, survival analysis is performed with one mutation, it is clear that there can be other mutations in HERVs which can have a profound impact on cancer progression. The ultimate goal of this study is to provide directions and suggestions for further research related to deciphering the role of HERVs in cancer.

Conclusion

In this study, we explored the correlation between HERV elements on human protein coding and non-coding regions and multiple cancers based on SNV hotspots. In the HERV element on human protein coding regions, we found four HERV elements that had over-represented nsSNVs and also overlapped with exons of four genes. Additionally, these four HERV elements were associated with at least 14 cancer types—notably skin and lung cancer. We showed that kidney cancer patients with the specific amino acid mutation A757G in ZNF99 within the HERV-W/LTR17/HERV17 element had a lower survival rate based on a survival analysis. We believe that this key mutation could play an important role in kidney cancer.

In the HERV element on human protein non-coding regions, we found 357 canonical and 431 non-canonical HERV elements across different classes which had significantly elevated SNVs counts. All SNVs within these 788 HERV elements overlapped with six DNA functional element groups. HERVs involved in the functional groups lncRNA, introns, and TFBS were shown to be associated with skin, esophageal, and liver cancer. Since we were able to narrow the number of cancer-related SNVs within HERV elements into six groups, we believe these are high-priority experimental targets for studying the molecular mechanisms in cancer progression.
Supporting information

S1 File. a–x The distribution of nsSNVs in protein coding region and SNVs in non-coding region by chromosome 1–22, X, Y. X-axis indicates the genomic positions of chromosome. Y-axis indicates the count of mutations. This Manhattan plot represents the distribution of mutations in all positions of the chromosome except for the centromere (empty region).

S1 Fig. Each HERV element with count of somatic SNVs within non-coding region. X-axis indicates cancer types. Y-axis indicates Log2Counts (counts mean the number of SNVs). Box plot represents the number of SNVs located in the HERV elements.

S1 Table. HERV elements dataset.

S2 Table. List of 39 canonical and 31 non-canonical HERV clades found in GRCh37/hg19.

S3 Table. The number of SNVs in every 1000 bases and their observed ratio in 22+XY chromosomes.

S4 Table. HERV elements on human protein coding region with significant representation of nsSNVs.

S5 Table. HERV elements on human non-coding region with significant representation of SNVs.

S6 Table. Somatic nsSNVs in HERV elements on human protein coding region.

S7 Table. Somatic SNVs in HERV elements on human non-coding region.

S8 Table. The regions of exon in four genes overlap to HERV elements.

S9 Table. 492 nsSNVs in HERV elements on human protein coding region with significant over-representation.

S10 Table. 193,439 SNVs in HERV elements on human non-coding region with significant over-representation.

S11 Table. The total count of nsSNVs in functional sites within HERV elements.

S12 Table. SNVs in HERV elements on human non-coding region overlap to DNA functional elements.
Acknowledgments
The authors would also like to thank Hiral Desai for review and critical feedback.

Author Contributions

**Conceptualization:** Ting-Chia Chang, Santosh Goud, John Torcivia-Rodriguez, Jonas Blomberg, Raja Mazumder.

**Data curation:** Ting-Chia Chang, Santosh Goud, John Torcivia-Rodriguez, Yu Hu.

**Formal analysis:** Ting-Chia Chang, Santosh Goud, John Torcivia-Rodriguez, Yu Hu, Qing Pan, Robel Kahsay.

**Funding acquisition:** Raja Mazumder.

**Investigation:** Ting-Chia Chang, Santosh Goud, John Torcivia-Rodriguez.

**Methodology:** Ting-Chia Chang, Santosh Goud, John Torcivia-Rodriguez, Yu Hu, Qing Pan, Raja Mazumder.

**Resources:** Ting-Chia Chang, Santosh Goud, John Torcivia-Rodriguez, Jonas Blomberg, Raja Mazumder.

**Software:** Ting-Chia Chang, John Torcivia-Rodriguez, Robel Kahsay.

**Supervision:** Robel Kahsay, Jonas Blomberg, Raja Mazumder.

**Validation:** Ting-Chia Chang, Santosh Goud, John Torcivia-Rodriguez, Yu Hu, Qing Pan, Jonas Blomberg, Raja Mazumder.

**Visualization:** Ting-Chia Chang, Yu Hu.

**Writing – original draft:** Ting-Chia Chang, Santosh Goud, Qing Pan.

**Writing – review & editing:** Ting-Chia Chang, Santosh Goud, John Torcivia-Rodriguez, Yu Hu, Jonas Blomberg, Raja Mazumder.

References


34. Lehrach H. DNA sequencing methods in human genetics and disease research. F1000prime reports. 2013; 5:34. https://doi.org/10.12703/P5-34 PMID: 24049638; PubMed Central PMCID: PMC3768324.


SNVs in human endogenous retrovirus and their association with cancer


