cancers

MDPI

Article

Transcriptome Analysis of Human Endogenous Retroviruses at Locus-Specific Resolution in Non-Small Cell Lung Cancer

Alessandro La Ferlita [1,*,†], Rosario Distefano [1,†], Salvatore Alaimo [2], Joal D. Beane [3], Alfredo Ferro [2], Carlo M. Croce [1], Philip N. Tsichlis [1], Alfredo Pulvirenti [2,‡] and Giovanni Nigita [1,*,‡]

1 Department of Cancer Biology and Genetics, The James Comphrensive Cancer Center, The Ohio State University, Columbus, OH 43210, USA
2 Bioinformatics Unit, Department of Clinical and Experimental Medicine, University of Catania, 95125 Catania, Italy
3 Department of Surgery, Division of Surgical Oncology, The James Comphrensive Cancer Center, The Ohio State University, Columbus, OH 43210, USA
* Correspondence: alessandro.laferlita@osumc.edu (A.L.F.); giovanni.nigita@osumc.edu (G.N.)
† These authors share equal contributions.
‡ These authors share equal senior authorship.

Simple Summary: Lung cancer is the leading cause of cancer deaths worldwide. Most lung cancer patients are diagnosed with locally advanced or metastatic diseases, and their prognosis is relatively poor, with 5-year survival rates ranging from 4 to 17%. Consequently, the identification of novel diagnostic lung cancer biomarkers remains crucial. Recently, human endogenous retroviruses (HERVs) have been found to be implicated in cancer development and later employed as novel diagnostic and prognostic cancer biomarkers. In this study, we present the first-ever locus-specific analysis of HERV expression in 515 lung adenocarcinoma (LUAD) and 497 lung squamous cell carcinoma (LUSC) patients' samples from the TCGA repository. In our study, we identified the differentially expressed HERVs in both TCGA-LUAD and TCGA-LUSC cohorts, we examined their impact on signaling pathways using in silico models, and we described HERVs' association with overall survival (OS) and relapse-free survival (RFS).

Abstract: Lung cancer is the second most commonly diagnosed cancer and the leading cause of cancer deaths worldwide. Among its subtypes, lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) are the most common, accounting for more than 85% of lung cancer diagnoses. Despite the incredible efforts and recent advances in lung cancer treatments, patients affected by this condition still have a poor prognosis. Therefore, novel diagnostic biomarkers are needed. Recently, a class of transposable elements called human endogenous retroviruses (HERVs) has been found to be implicated in cancer development and later employed as novel biomarkers for several tumor types. In this study, we first ever characterized the expression of HERVs at genomic locus-specific resolution in both LUAD and LUSC cohorts available in The Cancer Genome Atlas (TCGA). Precisely, (i) we profiled the expression of HERVs in TCGA-LUAD and TCGA-LUSC cohorts; (ii) we identified the dysregulated HERVs in both lung cancer subtypes; (iii) we evaluated the impact of the dysregulated HERVs on signaling pathways using neural network-based predictions; and (iv) we assessed their association with overall survival (OS) and relapse-free survival (RFS). In conclusion, we believe this study may help elucidate another layer of dysregulation that occurs in lung cancer involving HERVs, paving the way for identifying novel lung cancer biomarkers.

Keywords: human endogenous retroviruses; HERVs; lung cancer; lung adenocarcinoma; lung squamous cell carcinoma; biomarkers; RNA-seq; transcriptome analysis

Citation: La Ferlita, A.; Distefano, R.; Alaimo, S.; Beane, J.D.; Ferro, A.; Croce, C.M.; Tsichlis, P.N.; Pulvirenti, A.; Nigita, G. Transcriptome Analysis of Human Endogenous Retroviruses at Locus-Specific Resolution in Non-Small Cell Lung Cancer. *Cancers* **2022**, *14*, 4433. https://doi.org/10.3390/cancers14184433

Academic Editors: Amyn M. Rojiani, Srikumar Chellappan and Mumtaz Rojiani

Received: 13 July 2022
Accepted: 10 September 2022
Published: 13 September 2022

1. Introduction

Lung cancer, a malignancy that originates in the epithelium of the respiratory tracts, is the second most commonly diagnosed cancer and the leading cause of cancer deaths worldwide [1,2]. Traditionally, lung cancer is divided into two main groups: small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC) [3]. The NSCLC represents the most common lung cancer form, which comprises three histological subtypes: adenocarcinoma (LUAD) (~50%), squamous cell carcinoma (LUSC) (~35%), and large cell carcinoma (~15%) [4], with LUAD and LUSC accounting for more than 85% of NSCLC diagnosis worldwide. A lung cancer diagnosis is mainly carried out through imaging technologies and pathological examination. Both strategies present limitations in terms of sensitivity, with only 10–15% of new cases being diagnosed at an early clinical stage [5–7]. Hence, most lung cancer patients are diagnosed with locally advanced or metastatic diseases, and their prognosis is relatively poor, with 5-year survival rates ranging from 4 to 17% depending on the stage and metastasis location [8,9]. Consequently, the identification of novel diagnostic biomarkers remains crucial.

Recently, human endogenous retroviruses (HERVs), a class of transposable elements, have been found to be implicated in cancer development and later employed as novel diagnostic, prognostic, and treatment response biomarkers for several tumor types [10–15]. In more detail, HERVs are remnants of exogenous retroviruses integrated into the human genome during evolution, accounting for 8% of the human genome. These endogenized forms of viral sequences become established after germ cell infections by exogenous retroviruses. Once successfully integrated into the germline genome, proviruses are transmitted vertically by standard Mendelian inheritance [16]. Indeed, HERVs share genomic similarities to other exogenous retroviruses whose genomes typically consist of a common set of at least four genes: (i) *gag*, which encodes structural matrix and capsid proteins; (ii) *pro*, which encodes the viral protease; (iii) *pol*, which encodes the retroviral enzymes reverse transcriptase and integrase; and (iv) *env*, encoding the viral envelope glycoproteins. Due to accumulated mutations, HERVs have preserved the features of their original provirus to a highly variable extent, ranging from the retention of a complete set of long terminal repeats (LTRs) and retroviral genes to the retention of only fragments of the parental viral genomes [17].

Based on their sequences, HERVs are usually classified into 12 major families [18]. Precisely, their nomenclature refers to the single-letter code of the amino acid carried by the tRNA that binds the tRNA binding site of the viral RNA of that given family used initially as a primer to start the reverse transcription. Among them, the HERV-K family, which harbors the lysine tRNA binding site, is the most recently acquired by humans around three million years ago, and it is commonly subdivided into 11 subgroups (HML-1 through HML-11) [18]. Due to their relatively recent integration into the human genome, members of the HERV-K subgroup, called HML-2, still contain genes with intact open reading frame (ORF) that can encode for retroviral proteins [10]. Moreover, the expression of such elements may be deregulated in several human cancers. However, the exact roles of these genes in cancer development and progression remain unknown [10]. For this reason, HERV-K HML-2 elements are attracting significant interest in cancer research [10]. However, it is not only the HERV-K elements that might be relevant for cancer development. Emerging evidence has shown that several families of HERVs with highly mutated inactive protein-encoding genes are still actively transcribed, producing HERV-derived non-coding RNAs (ncRNAs), whose function and impact on cancer development remain to be elucidated [16]. In addition, the LTRs present at the 5′ and 3′ extremities of the HERVs may recruit transcription factors enhancing the transcription of host cell genes located in their proximity which may, in turn, affect the gene expression and, therefore, the biology of the cancer cells [19]. Recent studies have also shown that HERVs may regulate the host immune response to cancer cells by several mechanisms: (i) by inducing a mimetic state of a viral infection; (ii) by generating tumor-specific antigens; or (iii) by inducing the expression of genes associated

with immune response [18]. Consequently, their potential importance in immunotherapy raised significant interest in HERVs in recent years [18].

Although important cancer initiatives such as The Cancer Genome Atlas (TCGA) [20] have released an enormous amount of RNA-seq data potentially useful for the characterization of HERV expression in several human cancers, very few studies [21–27] have examined HERV expression in the TCGA cohorts. In addition, some of these studies [24–27] leveraged pipelines not explicitly designed for HERV characterization and quantification, lacking the resolving power to quantify HERVs at the genomic locus level. Indeed, quantifying HERVs with standard short-read RNA-seq technologies is challenging due to the repetitive nature of such elements and the consequent uncertainty in fragment assignment because of sequence similarity. Therefore, specific bioinformatics pipelines must be used to address this task. Several approaches have been proposed that account for read mapping uncertainty using statistical models. Among them, Telescope [28] accurately estimates HERV expression at the locus-specific level. Telescope addresses uncertainty in fragment assignment by reassigning ambiguously mapped fragments to the most probable source transcript leveraging a Bayesian statistical model [28]. Benchmark analyses performed by the Telescope's authors showed that their tool outperforms other methods for HERV quantification, providing the highest resolution since it estimates their expression at the genomic locus level rather than at the HERV subfamily level [28]. Recently, Telescope was also successfully used to identify dysregulated HERVs in head-neck, prostate, breast, colon, and uveal melanoma cohorts from TCGA [21–23]. However, other common tumor types such as lung cancer are surprisingly still overlooked. In the present study, we developed an analysis workflow that relies on Telescope [28] for HERV quantification and analysis. For the first time, we performed a locus-specific characterization of the HERV expression in TCGA-LUAD and TCGA-LUSC cohorts, identifying the potential consequences of their deregulation in NSCLC. Specifically, (i) we profiled the expression of HERVs in TCGA-LUAD and TCGA-LUSC cohorts; (ii) we identified the dysregulated HERVs in both lung cancer types; (iii) we evaluated the impact of the dysregulated HERVs on signaling pathways using neural network-based predictions; and (iv) we assessed the association of HERV expression with overall survival (OS) and relapse-free survival (RFS).

2. Materials and Methods

2.1. Pre-Processing of the RNA-Seq Data

Paired-end RNA-seq data (Illumina) were downloaded in BAM format (.bam) from TCGA-LUAD and TCGA-LUSC cohorts using the Genomic Data Commons (GDC) data transfer tool [29]. We tested the read strandedness using the *how_are_we_stranded_here* Python library [30], retaining only unstranded samples. Of the two cohorts, only TCGA-LUAD presented 22 stranded samples out of 598 (576 retained), while all TCGA-LUSC samples (545) were unstranded. Only primary tumor (non-metastatic) and solid normal tissue (non-cancerous tissues adjacent to the tumor) samples were employed for downstream analyses. Downloaded BAM files were first converted to FASTQ format (.fq) using BioBamBam2 (bamtofastq function) and then trimmed using Trim Galore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/ (accessed on 29 April 2022)), which includes FASTQC [31] and Cutadapt [32]. Trimmed reads were then remapped to the human genome (HG38 assembly) using Bowtie2 [33], allowing up to 100 alignments per read (-k 100). Afterward, the mapped reads in SAM format were converted into BAM format, sorted (for coordinates), and indexed using Samtools [34]. Finally, HERV quantification was performed using Telescope [28] with its GTF annotation file. All the pre-processing analyses were performed using the Ohio supercomputer (OSC). A schematic representation of the data pre-processing together with the complete analysis workflow is shown in Figure S1. The generated raw read count matrices were then used as input for all the downstream analyses presented in this study (see following sections). Raw count matrices of HERVs across all analyzed samples for TCGA-LUAD and TCGA-LUSC cohorts can be found in Supplementary Files S1 and S2, respectively.

2.2. Differential HERV Expression Analysis

To perform the differential HERV expression (DE) analysis, we first normalized the raw read counts via trimmed mean of M values (TMM) using edgeR [35] and then filtered out low expressed HERVs, whose mean value was less than five across all samples. Afterward, the raw read counts of the retained HERVs was log2-transformed leveraging *Voom* and then used for the DE analysis, leveraging the Limma R package [36]. Two different DE analyses were performed. In the first DE analysis, we considered only those tumor samples with matched adjacent non-cancerous tissues taken from the same patient (paired analysis). In particular, we compared 58 tumor/normal samples in TCGA-LUAD and 47 tumor/normal samples in TCGA-LUSC. Instead, in the second DE analysis, we compared the tumor samples at the early stages (IA) vs. the tumor samples at the advanced stages (III and IV) for both TCGA-LUAD and TCGA-LUSC. Notably, not all tumor samples available for both TCGA-LUAD and TCGA-LUSC cohorts have information about their clinical stages. Therefore, only the tumor samples whose stage (IA, III, or IV) was reported were taken into consideration (244 Samples for TCGA-LUAD and 171 samples for TCGA-LUSC). In both analyses, only HERVs with a $|Log_2FC| > 0.58$ ($|Linear\ FC| > 1.5$) and an adjusted p-value (Benjamini–Hochberg correction) < 0.05 were considered differentially expressed. A schematic representation of the DE analysis and the complete analysis workflow is shown in Figure S1. The plots generated for showing the results of the differential HERV expression analysis, such as volcano plots, Venn diagrams, heatmaps, and circos plots, have been drawn using EnhancedVolcano [37], InteractiVenn [38], pheatmap (https://cran.r-project.org/web/packages/pheatmap/index.html (accessed on 16 May 2022)), and circlize [39], respectively.

2.3. Differential Gene Expression Analysis

The gene raw read counts (coding and non-coding RNAs) for the TCGA-LUAD and TCGA-LUSC cohorts were downloaded via the GDC data portal (TCGA v33). Only the tumor samples paired with the adjacent normal tissue samples were used for the differential gene expression analysis. We first normalized the gene raw read counts via TMM and then filtered out low expressed genes whose geometric mean was <1 TMM across all samples. Afterward, the raw read counts of the retained genes were log2-transformed leveraging *Voom* and used for the DE analysis using Limma. Genes with an adjusted p-value (Benjamini–Hochberg correction) < 0.05 were considered differentially expressed.

2.4. Correlation Analyses

The differentially expressed HERVs identified from the TCGA-LUAD and TCGA-LUSC cohorts were correlated with the differentially expressed genes (DEGs) in the same TCGA cohorts and also with a list of 13 relevant oncogenes and adaptive immunity regulators used as therapeutic targets for NSCLC treatment [40,41], focusing only on the primary tumor. Normal and metastatic samples were not considered for this analysis. HERV and gene raw read count matrices were normalized using the read per million (RPM) formula to scale the raw library sizes. Subsequently, we correlated the HERV expression (Spearman correlation) first with the DEGs and second with the list of selected oncogenes and adaptive immunity regulators for each cohort. The resulting p-values were adjusted using the Benjamini–Hochberg correction. Only DEGs with an absolute value of Spearman correlation coefficient > 0.2 and an adjusted *p*-value < 0.05 were considered correlated with that specific HERV, while only HERVs correlated with the selected oncogenes and adaptive immunity regulators with an absolute value of Spearman correlation coefficient > 0.35 and an adjusted *p*-value < 0.05 were plotted on the correlation networks generated by Cytoscape (v3.9.1) [42] for both lung cancer cohorts.

2.5. Pathway Analysis

To predict the effects of the dysregulated HERVs identified in the TCGA-LUAD and TCGA-LUSC cohorts on metabolic and signaling pathways, we performed a neural

network-based topological pathway analysis using MITHrIL [43]. For each differentially expressed HERV, we used the list of its correlated DEGs (identified as described in the previous paragraph) with their Entrez ids and respective Log_2FC values as an input for MITHrIL. Only DEGs with an absolute value of Spearman correlation coefficient > 0.2 and adjusted *p*-value (Benjamini–Hochberg correction) < 0.05 were considered correlated for that specific HERV and used for the HERV-specific pathway analysis. Hence, single MITHrIL analyses were run for each differentially expressed HERV identified either in TCGA-LUAD or TCGA-LUSC cohorts, and only pathways with a *p*-value < 0.05 were selected for the further clustering analysis. For the latter, we considered each pathway's "Corrected Accumulator" value generated by MITHrIL. First, a consensus clustering analysis was performed to identify the number of potential clusters for both HERVs and enriched pathways using the ConsensusClusterPlus R package [44]. Afterward, a heatmap was generated using the ComplexHeatmap R package [45], where the rows were the HERVs, and the columns were the enriched pathways. The Spearman correlation distance was employed for both columns and rows. Pathways that were enriched in less than 30% of the differentially expressed HERVs were excluded from the clustering analysis.

2.6. Survival Analysis

For the survival analysis we considered all those HERVs with a geometric mean > 1 RPM across all samples. After the filtering step, we normalized the expressed HERVs with the TMM method, and we finally carried out both OS and RFS analyses by employing a univariate Cox regression model leveraging the *CoxPHFitter* function from the *lifelines* (v0.27.0) Python package. HERVs with a BH-adjusted Cox *p*-value < 0.1 were considered significant.

3. Results

3.1. HERV Profiling in TCGA-LUAD and TCGA-LUSC

To investigate the HERV expression patterns in NSCLC, we analyzed RNA-seq data from the LUAD and LUSC cohorts available in the TCGA repository. In particular, we examined RNA-seq data from 515 and 497 primary tumor tissues for TCGA-LUAD and TCGA-LUSC cohorts, respectively. Normal samples available for both cohorts were not used in this analysis but used in the differential HERV expression analysis (see the following section). A summary of the available clinical information for both TCGA-LUAD and TCGA-LUSC cohorts is reported in Supplementary File S3.

We developed an in-house workflow (described in the Section 2 and Figure S1) that relies on Telescope [28] to accurately detect and quantify HERVs at the genomic locus. Of the 14,174 HERV genomic loci analyzed, 4227 (29.8%) and 4439 (31.3%) were found expressed (average RPM > 5) in TCGA-LUAD and TCGA-LUSC cohorts, respectively. Notably, the two cohorts shared a large set of expressed HERVs (73.3%) (Figure 1A). Across them, the HERV families HERV-H, HERV-K, and MER4 were among the most represented ones. Although these families include more elements than others, they still show a considerably higher percentage of expressed HERVs per family. The lists of the top 10 most represented HERV families for both TCGA-LUAD and TCGA-LUSC are shown in Table S1.

Afterward, we performed a uniform manifold approximation and projection (UMAP) map reduction analysis (https://cran.r-project.org/web/packages/umap/index.html (accessed on 11 May 2022)) of the top 100 most variable HERVs based on their Median Absolute Deviation (MAD) values. The results showed that LUAD and LUSC samples form two well-separated clusters indicating that these two different histological lung cancer subtypes have different HERV expression patterns (Figure 1B). Moreover, additional UMAP analyses using the top 100 most variable HERVs identified either in LUAD or in LUSC were performed to establish whether tumors of the same molecular subtype cluster together using the HERV expression. The molecular subtype classification for TCGA-LUAD and TCGA-LUSC was retrieved from Chen F. et al. [46], and it takes into consideration the expression and genetic alterations of several genes such as SOX2, PTEN, TP53, KRAS, EGFR, and many others. The analysis showed in LUAD three clusters where tumor samples

belonging to the molecular subtypes AD.1, AD.2, AD.3, AD.4, and AD.5a tend to cluster with members of the same molecular subtypes while the other subtype AD.5b present tumor samples that are more heterogeneous in terms of HERV expression (Figure 1C). On the other hand, in LUSC we saw two better-defined clusters, wherein one of these clusters, tumor samples belonging to the molecular subtypes SQ.1 were more represented, while in the other one the molecular subtypes SQ.2a and SQ.2b were predominant (Figure 1D). To see details of the molecular characteristics of each molecular subtype refer to the paper of Chen F. et al. [46].

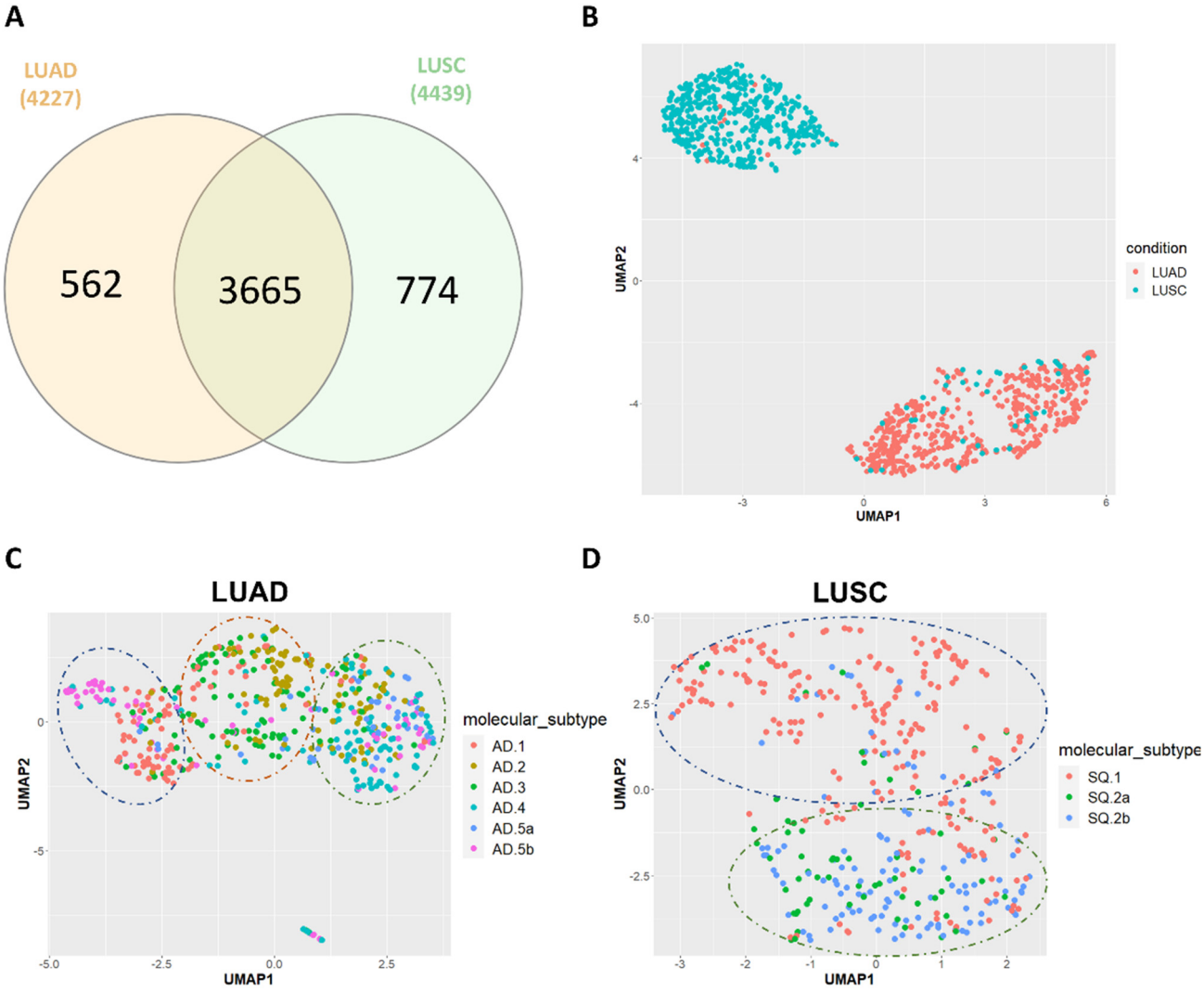

Figure 1. (**A**) Venn diagram showing the number of expressed HERVs identified exclusively in TCGA−LUAD, and TCGA−LUSC, and in common between two cohorts; (**B**) UMAP representation based on the top 100 most variable HERVs showing that TCGA−LUAD (red) and TCGA−LUSC (light blue) sample cohorts form two well-separated different clusters; UMAP representation based on the top 100 most variable HERVs in TCGA−LUAD (**C**) and TCGA−LUSC (**D**), showing the association between HERV expression and molecular subtypes.

3.2. Dysregulation of HERVs in TCGA-LUAD and TCGA-LUSC

To identify dysregulated HERVs in TCGA-LUAD and TCGA-LUSC cohorts, we performed differential expression analysis comparing the primary tumor samples (non-metastatic) with their adjacent normal tissue samples. For the TCGA-LUAD cohort, 58 tumor samples were compared with their respective 58 adjacent normal tissue samples, while for TCGA-LUSC, 47 tumor samples were compared with their respective 47 adjacent normal tissue samples.

Major dysregulations in HERV expression were observed in both cohorts. Specifically, a total of 1522 and 2421 HERVs were found dysregulated ($|Log_2FC| > 0.58$ and FDR < 0.05) in the TCGA-LUAD and TCGA-LUSC cohorts, respectively (Figure 2A,B). Dysregulated HERVs were found dispersed throughout the genome (Figure 2C,D).

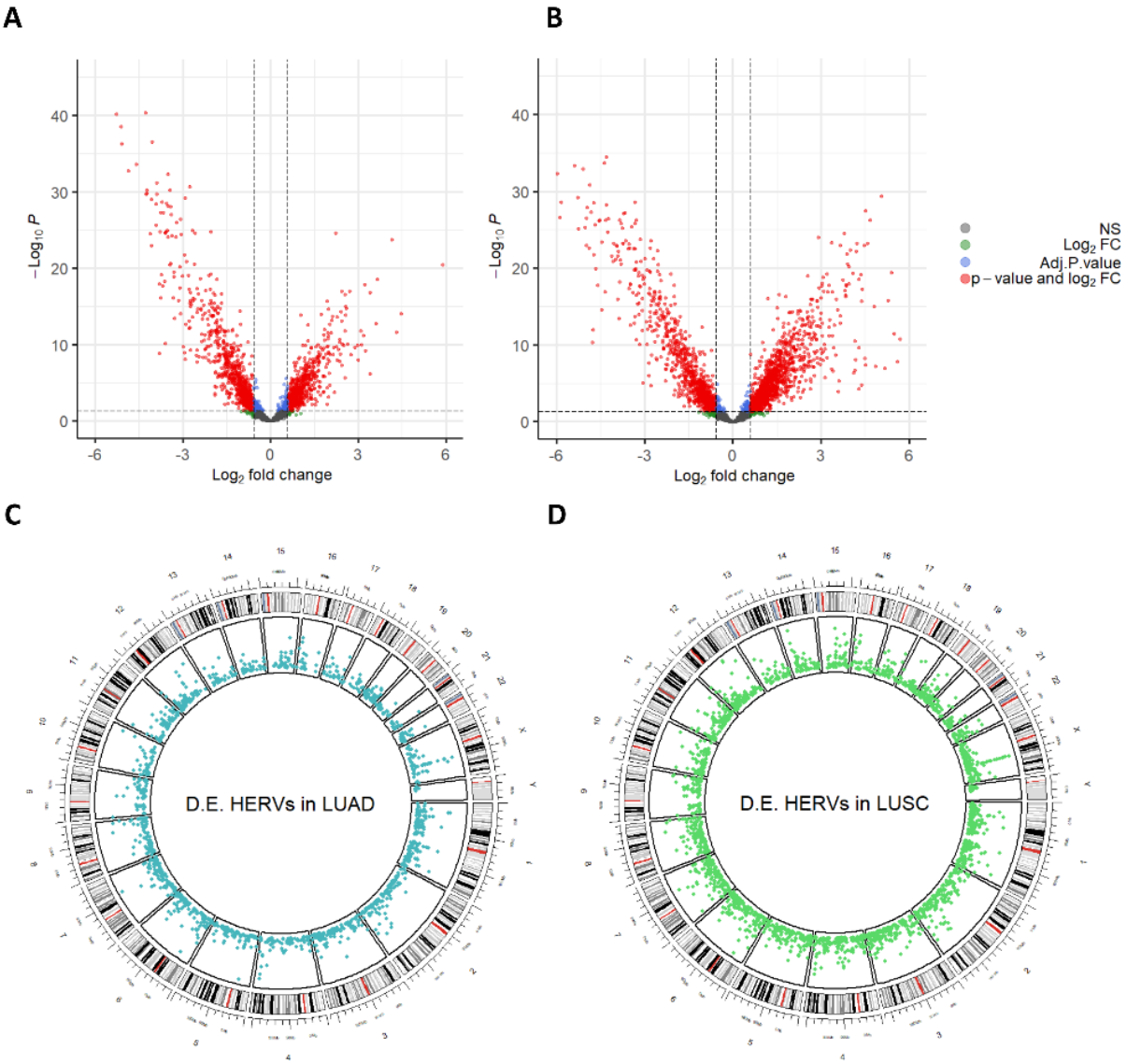

Figure 2. Volcano plots showing the distribution of the differentially expressed HERVs (red) identified in TCGA−LUAD (**A**) and TCGA−LUSC (**B**) in a two-dimensional space where the Y-axis is the significance ($-Log_{10}$P.value) and the X-axis is the magnitude of the change (Log_2FC); Circos plots showing the distribution of the differentially expressed HERVs identified in TCGA−LUAD (**C**) and TCGA−LUSC (**D**) across all human chromosomes.

We observed a slight predominance of down-regulated (820) HERVs over up-regulated (702) ones in TCGA-LUAD, while TCGA-LUSC showed an opposite trend (1435 up-regulated and 986 down-regulated). Only 27.6% of the up-regulated HERVs (Figure 3A) and 55.1% of the down-regulated (Figure 3B) HERVs overlapped between the two cohorts, suggesting different dysregulation mechanisms in the two lung cancer types. Furthermore, we built two heatmaps, one for TCGA-LUAD and one for TCGA-LUSC samples, using the normalized counts of the identified differentially expressed HERVs. The results for both cohorts revealed two well-defined clusters, one for tumors and one for normal samples

(Figure 3C,D). The differential HERV expression analysis results for LUAD and LUSC can be found in Supplementary File S4.

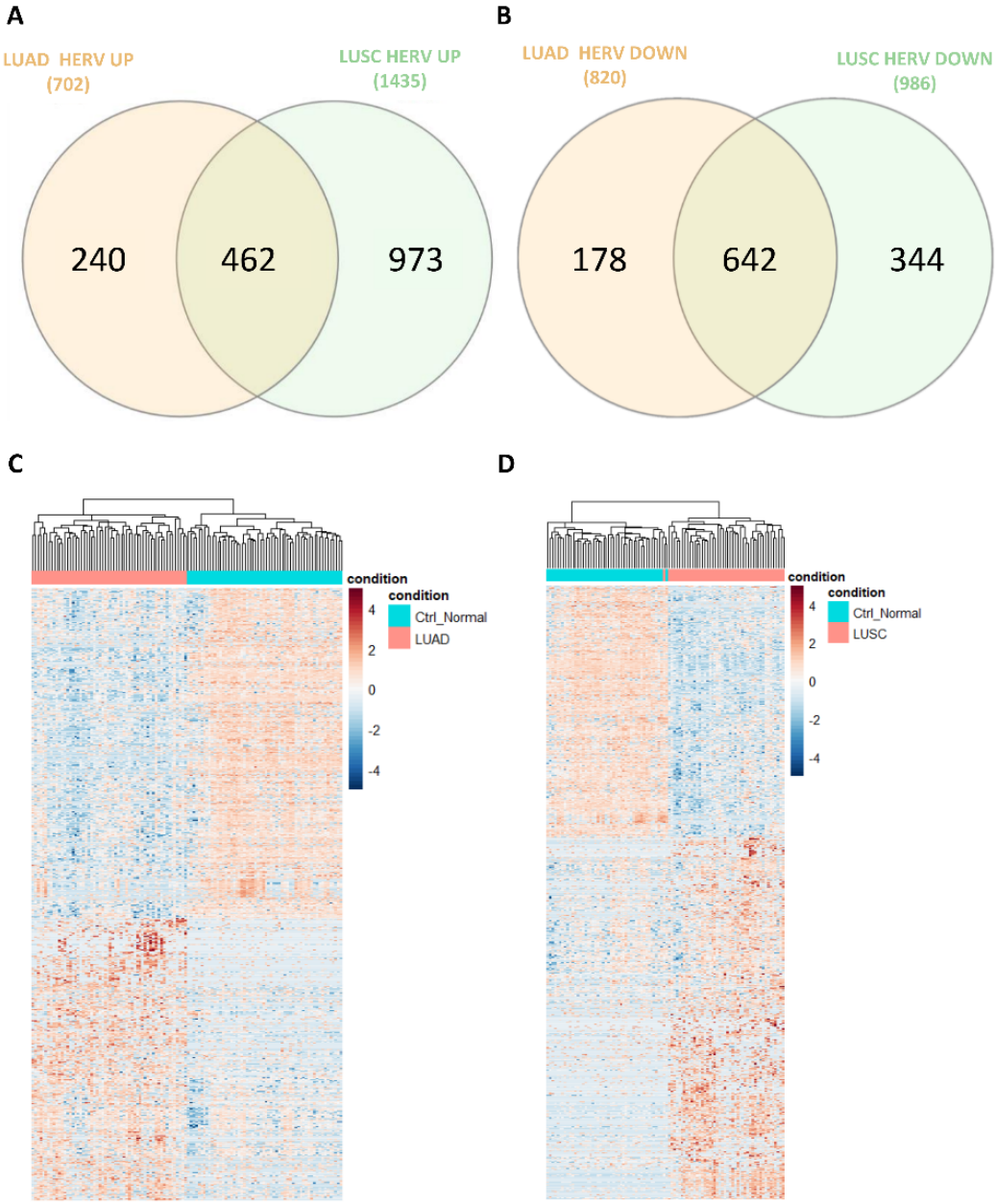

Figure 3. (**A**) Venn diagram showing the number of up-regulated HERVs identified exclusively in TCGA−LUAD, and TCGA−LUSC, and in common between these two cohorts; (**B**) Venn diagram showing the number of down-regulated HERVs identified exclusively in TCGA−LUAD, and TCGA−LUSC, and in common between these two cohorts; heatmap drawn by using the scaled counts (RPM) of the differentially expressed HERVs identified in TCGA−LUAD (**C**) and TCGA−LUSC (**D**) cohorts with their samples clustered accordingly with the Pearson correlation distance.

Moreover, several HERV families were found to be dysregulated in both LUAD and LUSC. Families with a higher percentage of differentially expressed HERVs included HERV-H, HERV-K, and MER4. Interestingly, up-regulation of the HERV-H and HERV-K families has been proposed as novel diagnostic biomarkers [47–49], while the up-regulation of the MER4 family has been linked to progression-free survival (PFS) and overall survival (OS) [11]. MER4 HERVs have also been proposed as novel biomarkers for the response to immune checkpoint inhibitors (ICI), which, currently, are becoming the standard first-line

treatment of advanced NSCLC for patients with high PD-L1 expression [11]. The lists of the top 10 HERV families dysregulated in LUAD and LUSC cohorts are shown in Table S2.

Finally, an additional differential expression analysis was performed comparing the early stages of NSCLC (IA) against the advanced stages (III and IV). The analysis, with the exception of TCGA-LUAD, did not show remarkable differences in HERV expression. Precisely, in TCGA-LUAD we identified 172 differentially expressed HERVs ($|Log2FC| > 0.58$ and $FDR < 0.05$), while in TCGA-LUSC, no strong significant results were detected. These results seem to suggest that the tumor stages clinically determined do not necessarily reflect a gradually increased dysregulation in HERV expression. The results of this analysis are reported in the Supplementary File S5.

3.3. HERV-Specific Neural Network-Based Pathway Analysis

To predict the effects of the HERV dysregulation on signaling pathways in the TCGA-LUAD and TCGA-LUSC cohorts, we employed MITHrIL [43] to perform a neural network-based topological pathway analysis focusing on genes whose expression correlates with HERV expression. First, we used a Spearman correlation formula to determine the correlation of the differentially expressed HERVs in either TCGA-LUAD or TCGA-LUSC cohorts with the DEGs identified in the respective cohort (see Materials and Methods section). The list of DEGs whose expression correlates with a given HERV was used as input for MITHrIL to identify the signaling pathways potentially impacted by the given HERV. Subsequently, clustering analysis based on the "Corrected Accumulator" values, which are an indicator of pathway activity generated by MITHrIL (a positive value indicates the up-regulation of a pathway while a negative value indicates the downregulation of a pathway), was performed to identify groups of HERVs that potentially impact pathways in a similar fashion (see Materials and Methods section). A schematic representation of the pathway analysis together with the complete analysis workflow is presented in Figure S1.

The results of the HERV-specific pathways analysis showed that several cancer-related signaling pathways were enriched in TCGA-LUAD and TCGA-LUSC cohorts (Figure 4A,B). Relevant examples include the MAPK signaling pathway, the mTOR signaling pathway, the RAS signaling pathway, the PI3K-Akt signaling pathway, the ErbB signaling pathway, the cGMP-PKG signaling pathway, the HIF-1 signaling pathway, and the cAMP signaling pathway (Figure 4A,B) suggesting that HERVs may be involved in important signaling pathways known as critical in cancer development and progression. Moreover, also several pathways that contribute to adaptive immunity were enriched in LUAD and LUSC cohorts. Relevant examples include the T cell receptor signaling pathway, the B cell receptor signaling pathway, and natural killer cell-mediated cytotoxicity (Figure 4A,B).

Importantly, the HERV profile correlations with enriched signaling pathways identified by MITHrIL in LUAD and LUSC exhibited significant differences. This observation reflects the fact that only 27.6% of up-regulated and 55.1% of down-regulated HERVs were in common between the two cohorts. The pathway analysis results reporting the corrected accumulators generated by MITHrIL for both TCGA-LUAD and TCGA-LUSC can be found in Supplementary File S6.

In conclusion, the clustering analysis on the HERV-specific pathway analysis results identified four main clusters of HERVs in TCGA-LUAD and three main clusters of HERVs in TCGA-LUSC, in which within the clusters, HERVs seem to impact the enriched pathways similarly (the list of HERVs for each of these clusters in both TCGA-LUAD and TCGA-LUSC can be found in Supplementary File S7). Moreover, in TCGA-LUAD, we observed a higher percentage of HERVs belonging to the HERV-H, HERV-K, and MER4 families in the fourth cluster compared to the other clusters, while in TCGA-LUSC, we observed a higher percentage of HERVs belonging to the HERV-H family in the second cluster, and HERV-K in the first cluster. On the other hand, the other HERV families were more similarly distributed along the clusters. Tables S3 and S4 show the top 10 dysregulated HERV-families for each cluster identified in TCGA-LUAD and TCGA-LUSC cohorts, respectively.

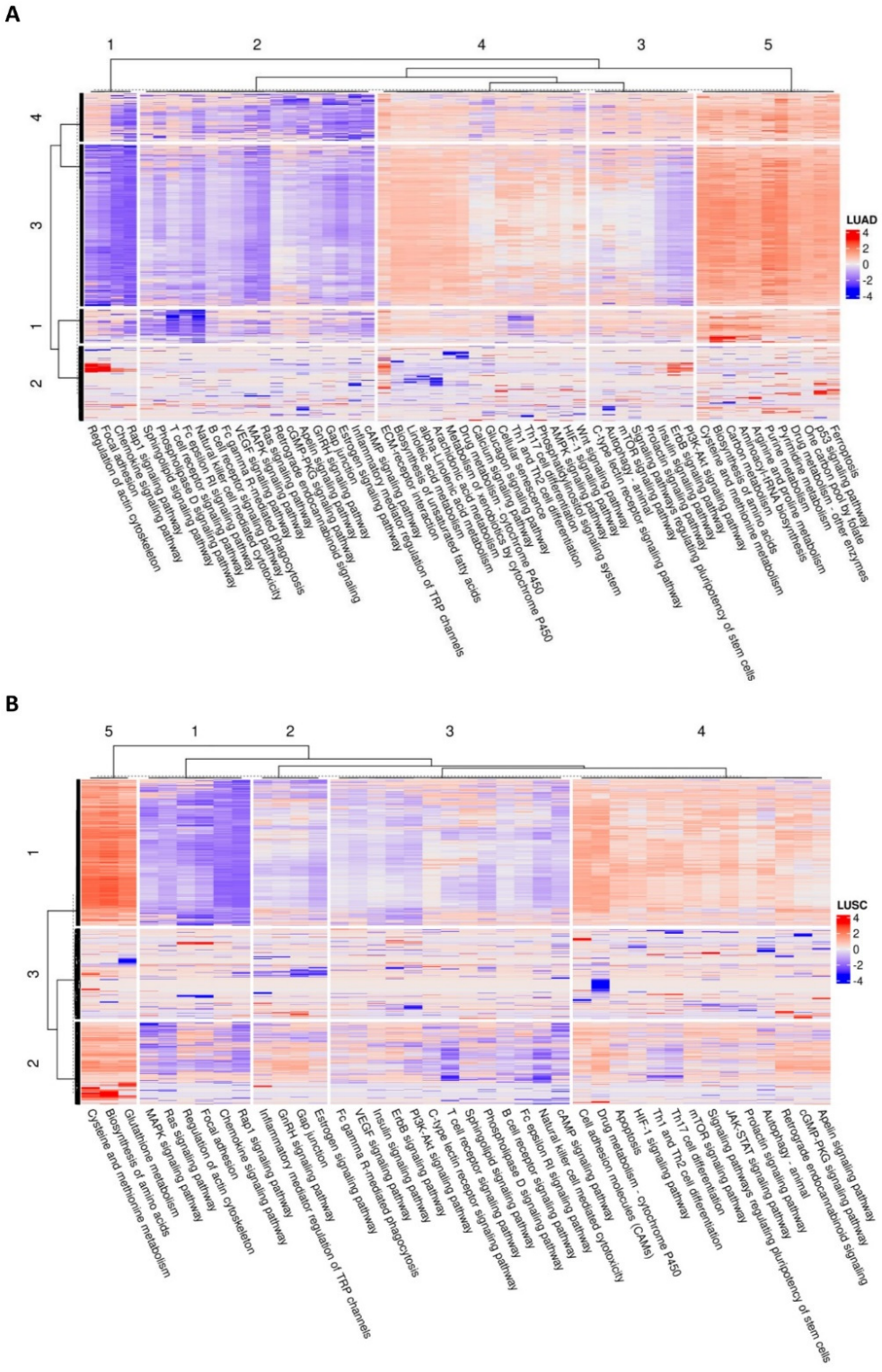

Figure 4. Heatmaps showing the most enriched pathways detected in TCGA−LUAD (**A**) and TCGA−LUSC (**B**) after the HERV−specific pathway analysis performed by MITHrIL. In more detail, the value reported in the heatmap are the corrected accumulators generated by MITHrIL when it computed the perturbation scores, and both columns (pathways) and rows (differentially expressed HERVs) were clustered using the Spearman correlation distance.

3.4. Correlation of HERVs with Oncogenes and Adaptive Immunity Regulators in TCGA-LUAD and TCGA-LUSC

The pathway analysis (described in the previous paragraph) showed enrichment of several signaling pathways notoriously known to be relevant in cancer progression and tumor immunity indicating a possible regulation of them by the HERVs identified as differentially expressed in TCGA-LUAD and TCGA-LUSC. To investigate more, we selected some genes involved in the regulation of such pathways in order to see their connections with the HERVs identified as differentially expressed. Precisely, we selected ten oncogenes and three adaptive immunity regulators used in the clinic as therapeutic targets for NSCLC treatment [40,41] and we built two correlation networks (one for LUAD and one for LUSC, respectively) showing their connections with the HERVs.

The analysis showed several HERVs to be statistically significantly correlated with the selected genes. In more detail, as can be observed from the correlation networks shown in Figure 5A,B for TCGA-LUAD and TCGA-LUSC, respectively, we identified many HERVs to be positively correlated with the selected genes (red edges) and very few negatively correlated (blue edges). For graphic reasons, only the HERVs highly correlated with the selected genes ($|\text{Spearman correlation coefficient}| > 0.35$ and $\text{FDR} < 0.05$) were plotted in the correlation networks (some of the selected genes are not reported in the correlation networks because they did not present correlated HERVs over the cutoff). Most of the genes presented a variable number of HERVs that exclusively correlated with their expression. Interestingly, ROS1 presented the biggest cluster for both TCGA-LUAD and TCGA-LUSC (Figure 5A,B), where 100 HERVs were exclusively correlated with its expression in LUAD, 90 HERVs in LUSC, of which 38 HERVs were in common between the two cohorts.

Notably, PDCD1 (PD-1), CD274 (PD-L1) and CTLA4, used as targets of immune-checkpoints inhibitors, had a considerable number of HERVs to be positively correlated with at least two of them (Figure 5A,B). Precisely, we identified 44 HERVs correlated with the aforementioned genes in LUAD, 66 HERVs in LUSC, of which 29 HERVs were in common between the two lung cancer histological subtypes. These results seem to suggest a possible involvement of these elements in the regulation of relevant immune signaling pathways and, potentially, the response to immune-checkpoints inhibitors.

Remarkably, even NTRK1, which has been recently associated to promote resistance to PD-1 inhibitors [50], showed several HERVs correlated with its expression that were also correlated with PDCD1 (PD-1), CD274 (PD-L1) expression in both LUAD and LUSC (Figure 5A,B) supporting the evidence that NTRK1 is involved in immune-checkpoints inhibitors response and that HERVs may play a role in orchestrating such phenomenon. The complete results of the correlation analyses can be found in the Supplementary File S8 while the correlation networks can be interactively explored by importing the Supplementary File S9 on Cytoscape [42].

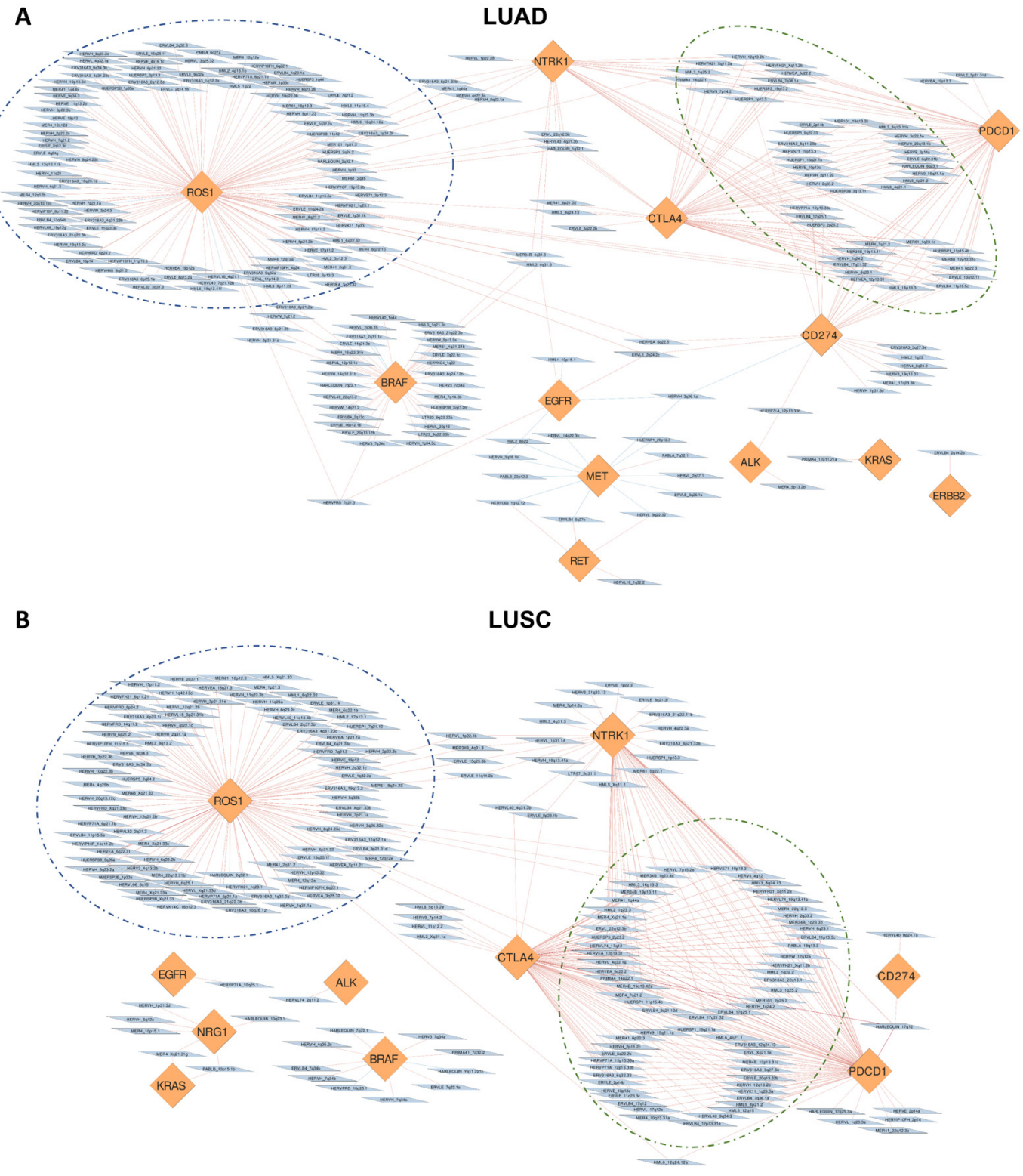

Figure 5. Correlation networks showing the connections between HERVs and a list of selected oncogenes and adaptive immunity regulators in TCGA−LUSC (**A**) and TCGA−LUSC (**B**). In the networks, the blue nodes represent the HERVs and the orange nodes represent the genes.

3.5. Association of HERVs with Survival in TCGA-LUAD and TCGA-LUSC

To assess the relationship between HERV expression and patients' survival in TCGA-LUAD and TCGA-LUSC cohorts, we conducted overall survival (OS) and relapse-free survival (RFS) analyses using a Cox proportional hazard regression model (see Section 2).

Interestingly, we identified 42 and 14 HERVs associated (Benjamini–Hochberg adjusted Cox *p*-value < 0.1) with poor OS and RFS in TCGA-LUAD samples, respectively (Figure 6). At the same time, we characterized 5 and 24 HERVs associated with poor OS and RFS in TCGA-LUSC samples, respectively (Figure 6).

LUAD OS

HERV	cox HR	cox FDR	D.E. status
ERVLE_14q32.32a	>1	1,43E-06	-
HERVL_11q22.3g	>1	3,88E-06	UP
ERV316A3_14q24.2f	>1	0,0006	-
HERVL74_10p11.22	>1	0,0012	-
HERVH_19q13.32e	>1	0,0026	-
ERVLB4_2q21.1c	>1	0,0033	UP
HERVH_7q31.1c	>1	0,0033	-
HERVL_2q37.1	>1	0,0041	UP
ERVLE_6q23.2b	>1	0,0043	-
HERVH_9p21.3a	>1	0,0123	UP
HERVH_11q14.3	>1	0,0144	UP
HERVFH19_9q34.11	>1	0,0152	UP
HERVH48_2q22.2	>1	0,0152	-
HERVH_14q22.1a	>1	0,0152	DOWN
HERVH_12q14.1a	>1	0,0208	-
ERVLE_1q23.1a	>1	0,0223	UP
HERVH_3q13.31a	>1	0,0278	-
HERVL_5q14.3e	>1	0,0278	-
LTR19_9q34.11	>1	0,0278	UP
PABLA_7q32.1	>1	0,0325	UP
HERVL_3p13b	>1	0,0375	UP
ERVLE_8q24.13f	>1	0,0388	-
HML2_4q32.3	>1	0,0388	-
HERVH_2p16.1c	>1	0,0403	DOWN
HUERSP2_1q32.1	>1	0,0403	-
HML3_19p12b	<1	0,0468	-
ERVLE_15q21.2	>1	0,0484	UP
ERVLE_8q24.3i	>1	0,0484	UP
HERVH_14q24.2b	>1	0,0484	-
MER4_11p15.4a	>1	0,0484	-
HERVH_4q22.1h	>1	0,0514	-
HERV9_12q22b	>1	0,0535	-
HERVL_11q12.1d	<1	0,0535	-
HERVH_13q33.3	>1	0,0537	DOWN
MER4_14q32.31a	>1	0,0537	-
HERV3_8q11.23	>1	0,0540	-
HERVIP10FH_2p14	<1	0,0585	UP
HERVH_10p12.1c	>1	0,0590	UP
HUERSP1_15q23b	>1	0,0662	DOWN
ERVLE_5q12.3a	<1	0,0684	DOWN
ERVLB4_5q12.3a	<1	0,0722	DOWN
MER41_3p14.3b	<1	0,0774	-

LUAD RFS

HERV	cox HR	cox FDR	D.E. status
ERVLE_8q24.3h	>1	0,0028	UP
ERV316A3_14q21.1b	>1	0,0388	-
ERVLE_9q22.31	>1	0,0388	-
ERVLE_8q24.3i	>1	0,0595	UP
ERV316A3_14q24.2f	>1	0,0673	-
HERVL18_2q23.1	>1	0,0782	UP
HERV9_11q22.3	>1	0,0923	-
HERVH_19q13.32e	>1	0,0923	-
HERVL18_4q22.1	>1	0,0923	-
MER41_7q22.3a	>1	0,0923	-
MER4_7q22.3	>1	0,0923	-
ERVLE_3p12.2h	>1	0,0991	-
HML2_4q32.3	>1	0,0991	-
PABLA_8p21.3b	>1	0,0991	-

LUSC OS

HERV	cox HR	cox FDR	D.E. status
HERVH_8q24.23c	>1	0,0063	DOWN
HERVH_7q31.33b	>1	0,0231	UP
HERVH_13q14.11b	>1	0,0335	-
HML5_19p12b	>1	0,0423	UP
HML5_8p23.2b	>1	0,0765	DOWN

LUSC RFS

HERV	cox HR	cox FDR	D.E. status
ERV316A3_7q33a	>1	0,0113	UP
HERVEA_5q35.3	>1	0,0113	UP
HERVH_1q41f	>1	0,0113	-
HERVW_21q21.1	>1	0,0113	UP
HERVH_12q21.2a	>1	0,0128	-
HUERSP2_12q21.1	>1	0,0128	UP
HML2_21q21.1	>1	0,0186	UP
ERVLE_5p13.3m	>1	0,0254	UP
HERVH_6q14.1c	>1	0,0267	UP
HERVFRD_19q13.41	>1	0,0269	UP
HERVH48_7q22.1	>1	0,0337	-
HERVIP10FH_10q23.33	>1	0,0350	UP
LTR25_11q24.3b	>1	0,0370	UP
ERVLB4_8q23.3c	>1	0,0712	UP
HERVIP10F_3p26.1	>1	0,0756	UP
ERVLB4_Xp21.1b	>1	0,0894	-
ERVLE_20q13.12c	>1	0,0894	-
ERVLE_5q23.3g	>1	0,0894	UP
HERVH_13q14.11b	>1	0,0894	-
HERVIP10FH_14q12	>1	0,0894	-
HERVL_2q37.1	>1	0,0894	-
HERVL_9q13a	>1	0,0894	-
HERVW_3p22.2	>1	0,0894	UP
ERV316A3_4q28.3k	>1	0,0923	DOWN

Figure 6. Tables showing the list of HERVs statistically associated with overall survival (OS) and relapse-free survival (RFS) in both TCGA−LUAD and TCGA−LUSC cohorts. In more detail, Cox hazard ratios (HR), and Cox false discovery rates (FDR) values are reported. In addition, it is also reported for each HERV if it has been found up−regulated (up) or down−regulated (down) during the paired tumor vs. normal differential HERV expression analysis.

Significantly, most (89.3% in LUAD and 100% in LUSC) of potential prognostic HERVs have a hazard ratio greater than 1, indicating that their high expression is associated with a poor prognosis. Moreover, 32% of the HERVs associated with a poor OS or RFS in TCGA-LUAD, and 55.2% of the HERVs associated with poor OS or RFS in TCGA-LUSC, were also up-regulated in LUAD and LUSC (compared to their adjacent normal tissues). However, this is the first time that these HERVs have been linked to cancer biology; therefore, no additional information is available in the literature about their role in cancer development.

4. Discussion

This study presents a locus-specific analysis of HERV expression in LUAD and LUSC patients' samples from the TCGA repository. Precisely, we identified the differentially expressed HERVs in both TCGA-LUAD and TCGA-LUSC cohorts, we examined their impact on cell signaling pathways using in silico models, and we assessed HERVs' association with OS and RFS.

The analysis of transposable elements such as HERVs using standard short-read RNA-seq technologies is challenging because of the repetitive nature of such elements, and the consequent ambiguity in assigning RNA-seq reads to the most probable locus.

This problem has been recently addressed with the development of novel methodologies. In particular, Telescope [28] emerged over the other existing methods by allowing an accurate estimation of HERV expression. Telescope uses a Bayesian statistical model to reassign ambiguously mapped fragments to the most probable source gene. Moreover, it estimates HERV expression at the genomic locus level rather than at the HERV subfamily level, providing a higher resolution power than the other methods [28]. Based on these considerations, we design an in-house workflow that relies on Telescope for quantifying HERV expression from the RNA-seq data of the TCGA-LUAD and TCGA-LUSC cohorts (Figure S1).

Our analysis addressed the genomic mapping of 14,174 HERV loci, of which 4227 (29.8%) were expressed in the TCGA-LUAD cohort, and 4439 (31.3%) were expressed in the TCGA-LUSC cohort. The HERV subfamilies expressed in a higher percentage in TCGA-LUAD and TCGA-LUSC were HERV-H, HERV-K, and MER4 (Table S1). Although 73% of HERVs were expressed in common in both lung cancer types (Figure 1A), a map reduction analysis of the top 100 most variable HERVs showed that TCGA-LUAD and TCGA-LUSC samples were well-separated in two distinct clusters (Figure 1B). Additional cohort-specific UMAP analyses also showed a certain correlation between HERV expression and LUAD and LUSC molecular subtypes (Figure 1C,D) defined by Chen F. et al. [46]. These take into consideration the expression and genetic alteration of several genes that are relevant in cancer biology [46].

Following the identification of HERVs expressed in TCGA-LUAD and TCGA-LUSC, we questioned which of those HERVs may be up-regulated or down-regulated in tumor samples relative to their adjacent normal tissues. This analysis identified 1522 HERVs dysregulated in TCGA-LUAD and 2421 HERVs dysregulated in TCGA-LUSC (Figure 2A,B). Dysregulated HERVs were distributed throughout the genome (Figure 2C,D). Of the up-regulated HERVs, only 27.6% were up-regulated in both TCGA-LUAD and TCGA-LUSC (Figure 3A); at the same time, of the down-regulated HERVs, only 55.1% were down-regulated in both cohorts (Figure 3B), suggesting that different dysregulation mechanisms involving HERV expression occur in these two different lung cancer types. However, the most frequently dysregulated HERV families, such as HERV-H, HERV-K, and MER4, were the same in both TCGA-LUAD and TCGA-LUSC (Table S2). Notably, the upregulation of the HERV-H and HERV-K families in lung cancer has been observed by others, and it has been suggested as a potential diagnostic biomarker [47–49]. The expression of members of the MER4 family has been associated with differences in progression-free and overall survival. At the same time, it has been proposed as a biomarker for the response to immune checkpoint inhibitors [11].

To determine whether HERV dysregulation and functional characteristics of NSCLC are linked, we first examined the correlation between the expression of individual deregulated HERVs and the expression of DEGs (only protein-coding ones) in both TCGA-LUAD and TCGA-LUSC cohorts. The list of DEGs with a significant Spearman correlation coefficient for a given HERV was then used as input for the MITHrIL functional pathway analysis. MITHrIL [43] is a neural network-based topological pathway analysis algorithm that fully exploits the topological information encoded by the KEGG's pathways [51] to compute perturbation scores. KEGG's pathways are modeled in MITHrIL as complex graphs where each node is a biological element (protein-coding gene, miRNA, or metabolite), and each edge is an interaction between nodes. The functional enrichment analysis identified several pathways notoriously known to be crucial in cancer development and progression such as MAPK, PI3K/AKT/mTOR, RAS, ErbB, and HIF-1 signaling pathways as well as pathways involved in adaptive immunity in both TCGA-LUAD and TCGA-LUSC cohorts (Figure 4A,B). Further correlation network analyses between HERV expression and key genes involved in the aforementioned pathways, which are also used as targets in the clinical practice for NSCLC treatment, revealed interesting clusters of HERVs exclusively or commonly correlated with them in both LUAD and LUSC. Significantly, PDCD1 (PD-1), CD274 (PD-L1), and CTLA4, used as targets of immune-checkpoints inhibitors, had a

considerable number of HERVs to be positively correlated with at least two of them in both LUAD and LUSC (Figure 5A,B) suggesting a possible involvement of these elements in the regulation of relevant immune signaling pathways and, potentially, the response to immune-checkpoints inhibitors. These results combined seem to point out that the dysregulation of these pathways in NSCLC may also be caused by the observed dysregulation in HERV expression. HERV expression may, in fact, affect tumor cell functions via multiple mechanisms. First, the LTRs of both complete and defective HERVs contain enhancer and promoter elements whose activity may profoundly affect the expression of neighboring genes. If we consider that at least 20% of the human genes are adjacent to HERVs, it becomes clear that changes in the activity of the HERV's regulatory elements may have profound effects on the biology of the tumor cells. Another mechanism is the expression of proteins, such as Np9 and Rec, which have been reported to interact with and regulate the activity of transcription factors that, in turn, may promote cancer progression [52–54]. In addition to these mechanisms that may promote cancer progression, others may have an opposite effect [55–57]. A relevant example includes the expression of viral envelope proteins encoded by intact or near intact members of the HERV-K family, particularly members of the HML-2 subtype, which act as neo-antigens, initiating an adaptive immune response against the tumor cells [10,55,58,59]. Although over the last years several studies have shown that HERV expression might activate an adaptive anti-tumor immunity in multiple human cancers, progress in this area is impeded by the lack of detailed information on the specific members of HERV families who are dysregulated in a given type of tumor. However, prominent immunotherapy clinical trial studies in lung cancer such as POPLAR [60] and OAK [61] have recently publicly released the RNA-seq data that can be potentially reanalyzed for assessing HERV expression, giving the opportunity for evaluating their importance in regulating adaptive immunity and affecting the response to immune-checkpoints inhibitors.

HERV expression-associated functional changes in NSCLCs may alter the NSCLC natural history. To address this question, we examined the association of HERV expression with OS and RFS in both TCGA-LUAD and TCGA-LUSC. This analysis identified 42 and 14 HERVs associated with poor OS and RFS in TCGA-LUAD samples, respectively (Figure 6). At the same time, we also identified 5 and 24 HERVs associated with poor OS and RFS in TCGA-LUSC samples, respectively (Figure 5). Significantly, the vast majority of these HERVs showed a hazard ratio greater than 1, indicating that their high expression is related to a worse prognosis. Moreover, 32% of the HERVs associated with a poor OS or RFS in TCGA-LUAD, and 55.2% of the HERVs associated with poor OS or RFS in TCGA-LUSC, were also up-regulated in LUAD and LUSC (compared to their adjacent normal tissues) suggesting a possible involvement of these HERVs in NSCLC progression. Although the link between HERV expression and patient survival may indicate a causative relationship, this connection may be caused by unknown mechanisms, where HERV expression should be viewed as a valuable biomarker. In this case, the cause of the association should be addressed in future studies.

5. Conclusions

This study represents the first locus-specific transcriptome analysis of HERVs in lung cancer. We described dysregulations in HERV expression in TCGA-LUAD and TCGA-LUSC; we also identified their potential impact on regulating essential signaling and immune system pathways, which may affect lung cancer development and progression. Finally, we reported several HERVs associated with worse OS and RFS in LUAD and LUSC. In conclusion, we believe this study might pave the way for identifying novel HERV-based biomarkers in lung cancer and suggests that such HERVs require further investigation and validation to reveal their involvement in crucial signaling pathways.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/cancers14184433/s1, Figure S1: Schematic presentation of the analysis workflow: from the RNA-seq data pre-processing to the downstream analyses; Table S1: Table reporting the top 10 most expressed HERV families in TCGA-LUAD and TCGA-LUSC cohorts; Table S2: Table reporting the top 10 most dysregulated HERV families found in TCGA-LUAD and TCGA-LUSC cohorts; Table S3: Table reporting the top 10 most dysregulated HERV-families for each cluster (four clusters in total) found during the clustering analysis of the HERV-specific pathway analysis in TCGA-LUAD; Table S4: Table reporting the top 10 most dysregulated HERV-families for each cluster (three clusters in total) found during the clustering analysis of the HERV-specific pathway analysis in TCGA-LUSC; File S1: Tab-separated values file (.tsv) reporting the raw counts (not-normalized) assigned by Telescope across all analyzed TCGA-LUAD samples; File S2: Tab-separated values file (.tsv) reporting the raw counts (not-normalized) assigned by Telescope across all analyzed TCGA-LUSC samples; File S3: Excel file reporting in two separate sheets the patients' clinical information about the TCGA-LUAD and TCGA-LUSC cohorts; File S4: Excel file reporting in two separate sheets the results of the differential HERV expression analysis performed from the paired tumor vs. normal samples for both TCGA-LUAD and TCGA-LUSC cohorts; File S5: Excel file reporting in two separate sheets the results of the differential HERV expression analysis comparing for both TCGA-LUAD and TCGA-LUSC cohorts the tumor samples at the early stages (IA) against the tumor samples at the advanced stages (III and IV); File S6: Excel file reporting in two separate sheets the corrected accumulators generated by MITHrIL pathway analysis for both TCGA-LUAD and TCGA-LUSC cohorts; File S7: Excel file reporting in two separate sheets the list of the dysregulated HERVs assigned for each cluster identified during the clustering analysis of the HERV-specific pathway analyses for both TCGA-LUAD and TCGA-LUSC cohorts; File S8: Excel file reporting in two separate sheets the results of the correlation analysis between the differentially expressed genes identified in TCGA-LUAD and TCGA-LUSC and the expression of the selected oncogenes and immune-checkpoints regulators in the same cohorts; File S9: Input file for Cytoscape containing the correlation networks generated for TCGA-LUAD and TCGA-LUSC.

Author Contributions: Conceptualization, A.L.F. and G.N.; methodology, A.L.F., R.D., S.A. and G.N.; software, A.L.F. and R.D.; formal analysis, A.L.F., R.D., S.A. and G.N.; data curation, A.L.F. and R.D.; writing—original draft preparation, A.L.F. and G.N.; writing—review and editing, A.L.F., R.D., S.A., J.D.B., A.F., C.M.C., P.N.T., A.P. and G.N.; visualization, A.L.F.; supervision, A.P. and G.N.; project administration, G.N.; funding acquisition, C.M.C., P.N.T. and G.N. All authors have read and agreed to the published version of the manuscript.

Funding: The work was supported by the National Cancer Institute grant P30 CA016672 to the Ohio State University Comprehensive Cancer Center.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Raw sequence data were retrieved via the GDC data portal after obtaining authorization from the data access committee (DBGap Project ID: 11332).

Acknowledgments: The results shown in this study are based on data generated by the TCGA Research Network (http://cancergenome.nih.gov/ (accessed on 29 April 2022)). We thank John Coffin for helpful scientific discussions. We want to thank the Cancer IT Operation Group of The Ohio State University, in particular, Thomas Moore, for his special technical assistance. We also want to thank the Ohio Supercomputer Center for its resources and technical support (Project ID: PDE0005).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Schwartz, A.G.; Cote, M.L. Epidemiology of Lung Cancer. *Adv. Exp. Med. Biol.* **2016**, *893*, 21–41.
2. Siegel, R.L.; Miller, K.D.; Jemal, A. Cancer Statistics, 2020. *CA Cancer J. Clin.* **2020**, *70*, 7–30. [CrossRef] [PubMed]
3. Howlader, N.; Forjaz, G.; Mooradian, M.J.; Meza, R.; Kong, C.Y.; Cronin, K.A.; Mariotto, A.B.; Lowy, D.R.; Feuer, E.J. The Effect of Advances in Lung-Cancer Treatment on Population Mortality. *N. Engl. J. Med.* **2020**, *383*, 640–649. [CrossRef] [PubMed]
4. Cagle, P.T.; Allen, T.C.; Olsen, R.J. Lung Cancer Biomarkers: Present Status and Future Developments. *Arch. Pathol. Lab. Med.* **2013**, *137*, 1191–1198. [CrossRef] [PubMed]

5. Patz, E.F., Jr.; Pinsky, P.; Gatsonis, C.; Sicks, J.D.; Kramer, B.S.; Tammemägi, M.C.; Chiles, C.; Black, W.C.; Aberle, D.R. NLST Overdiagnosis Manuscript Writing Team Overdiagnosis in Low-Dose Computed Tomography Screening for Lung Cancer. *JAMA Intern. Med.* **2014**, *174*, 269–274. [CrossRef]
6. Konopka, K.E. Diagnostic Pathology of Lung Cancer. *Semin. Respir. Crit. Care Med.* **2016**, *37*, 681–688. [CrossRef]
7. Xi, K.-X.; Zhang, X.-W.; Yu, X.-Y.; Wang, W.-D.; Xi, K.-X.; Chen, Y.-Q.; Wen, Y.-S.; Zhang, L.-J. The Role of Plasma miRNAs in the Diagnosis of Pulmonary Nodules. *J. Thorac. Dis.* **2018**, *10*, 4032–4041. [CrossRef]
8. Hirsch, F.R.; Scagliotti, G.V.; Mulshine, J.L.; Kwon, R.; Curran, W.J., Jr.; Wu, Y.-L.; Paz-Ares, L. Lung Cancer: Current Therapies and New Targeted Treatments. *Lancet* **2017**, *389*, 299–311. [CrossRef]
9. Santarpia, M.; Liguori, A.; D'Aveni, A.; Karachaliou, N.; Gonzalez-Cao, M.; Daffinà, M.G.; Lazzari, C.; Altavilla, G.; Rosell, R. Liquid Biopsy for Lung Cancer Early Detection. *J. Thorac. Dis.* **2018**, *10*, S882–S897. [CrossRef]
10. Curty, G.; Marston, J.L.; de Mulder Rougvie, M.; Leal, F.E.; Nixon, D.F.; Soares, M.A. Human Endogenous Retrovirus K in Cancer: A Potential Biomarker and Immunotherapeutic Target. *Viruses* **2020**, *12*, 726. [CrossRef]
11. Lecuelle, J.; Favier, L.; Fraisse, C.; Lagrange, A.; Kaderbhai, C.; Boidot, R.; Chevrier, S.; Joubert, P.; Routy, B.; Truntzer, C.; et al. MER4 Endogenous Retrovirus Correlated with Better Efficacy of Anti-PD1/PD-L1 Therapy in Non-Small Cell Lung Cancer. *J. Immunother. Cancer* **2022**, *10*, e004241. [CrossRef] [PubMed]
12. Golkaram, M.; Salmans, M.L.; Kaplan, S.; Vijayaraghavan, R.; Martins, M.; Khan, N.; Garbutt, C.; Wise, A.; Yao, J.; Casimiro, S.; et al. HERVs Establish a Distinct Molecular Subtype in Stage II/III Colorectal Cancer with Poor Outcome. *NPJ Genom Med* **2021**, *6*, 13. [CrossRef] [PubMed]
13. Tavakolian, S.; Goudarzi, H.; Faghihloo, E. Evaluating the Expression Level of HERV-K Env, np9, Rec and Gag in Breast Tissue. *Infect. Agent. Cancer* **2019**, *14*, 42. [CrossRef] [PubMed]
14. Wei, Y.; Wei, H.; Wei, Y.; Tan, A.; Chen, X.; Liao, X.; Xie, B.; Wei, X.; Li, L.; Liu, Z.; et al. Screening and Identification of Human Endogenous Retrovirus-K mRNAs for Breast Cancer Through Integrative Analysis of Multiple Datasets. *Front. Oncol.* **2022**, *12*, 820883. [CrossRef]
15. Manca, M.A.; Solinas, T.; Simula, E.R.; Noli, M.; Ruberto, S.; Madonia, M.; Sechi, L.A. HERV-K and HERV-H Env Proteins Induce a Humoral Response in Prostate Cancer Patients. *Pathogens* **2022**, *11*, 95. [CrossRef]
16. Grandi, N.; Tramontano, E. Human Endogenous Retroviruses Are Ancient Acquired Elements Still Shaping Innate Immune Responses. *Front. Immunol.* **2018**, *9*, 2039. [CrossRef]
17. Geis, F.K.; Goff, S.P. Silencing and Transcriptional Regulation of Endogenous Retroviruses: An Overview. *Viruses* **2020**, *12*, 884. [CrossRef]
18. Petrizzo, A.; Ragone, C.; Cavalluzzo, B.; Mauriello, A.; Manolio, C.; Tagliamonte, M.; Buonaguro, L. Human Endogenous Retrovirus Reactivation: Implications for Cancer Immunotherapy. *Cancers* **2021**, *13*, 1999. [CrossRef]
19. Gonzalez-Cao, M.; Iduma, P.; Karachaliou, N.; Santarpia, M.; Blanco, J.; Rosell, R. Human Endogenous Retroviruses and Cancer. *Cancer Biol. Med.* **2016**, *13*, 483–488.
20. Cancer Genome Atlas Research Network; Weinstein, J.N.; Collisson, E.A.; Mills, G.B.; Shaw, K.R.M.; Ozenberger, B.A.; Ellrott, K.; Shmulevich, I.; Sander, C.; Stuart, J.M. The Cancer Genome Atlas Pan-Cancer Analysis Project. *Nat. Genet.* **2013**, *45*, 1113–1120. [CrossRef]
21. Kolbe, A.R.; Bendall, M.L.; Pearson, A.T.; Paul, D.; Nixon, D.F.; Pérez-Losada, M.; Crandall, K.A. Human Endogenous Retrovirus Expression Is Associated with Head and Neck Cancer and Differential Survival. *Viruses* **2020**, *12*, 956. [CrossRef] [PubMed]
22. Steiner, M.C.; Marston, J.L.; Iñiguez, L.P.; Bendall, M.L.; Chiappinelli, K.B.; Nixon, D.F.; Crandall, K.A. Locus-Specific Characterization of Human Endogenous Retrovirus Expression in Prostate, Breast, and Colon Cancers. *Cancer Res.* **2021**, *81*, 3449–3460. [CrossRef] [PubMed]
23. Bendall, M.L.; Francis, J.H.; Shoushtari, A.N.; Nixon, D.F. Specific Human Endogenous Retroviruses Predict Metastatic Potential in Uveal Melanoma. *JCI Insight* **2022**, *7*. [CrossRef] [PubMed]
24. Johanning, G.L.; Malouf, G.G.; Zheng, X.; Esteva, F.J.; Weinstein, J.N.; Wang-Johanning, F.; Su, X. Expression of Human Endogenous Retrovirus-K Is Strongly Associated with the Basal-like Breast Cancer Phenotype. *Sci. Rep.* **2017**, *7*, 41960. [CrossRef] [PubMed]
25. Smith, C.C.; Beckermann, K.E.; Bortone, D.S.; De Cubas, A.A.; Bixby, L.M.; Lee, S.J.; Panda, A.; Ganesan, S.; Bhanot, G.; Wallen, E.M.; et al. Endogenous Retroviral Signatures Predict Immunotherapy Response in Clear Cell Renal Cell Carcinoma. *J. Clin. Investig.* **2018**, *128*, 4804–4820. [CrossRef]
26. Panda, A.; de Cubas, A.A.; Stein, M.; Riedlinger, G.; Kra, J.; Mayer, T.; Smith, C.C.; Vincent, B.G.; Serody, J.S.; Beckermann, K.E.; et al. Endogenous Retrovirus Expression Is Associated with Response to Immune Checkpoint Blockade in Clear Cell Renal Cell Carcinoma. *JCI Insight* **2018**, *3*. [CrossRef]
27. Kong, Y.; Rose, C.M.; Cass, A.A.; Williams, A.G.; Darwish, M.; Lianoglou, S.; Haverty, P.M.; Tong, A.-J.; Blanchette, C.; Albert, M.L.; et al. Transposable Element Expression in Tumors Is Associated with Immune Infiltration and Increased Antigenicity. *Nat. Commun.* **2019**, *10*, 5228. [CrossRef]
28. Bendall, M.L.; de Mulder, M.; Iñiguez, L.P.; Lecanda-Sánchez, A.; Pérez-Losada, M.; Ostrowski, M.A.; Jones, R.B.; Mulder, L.C.F.; Reyes-Terán, G.; Crandall, K.A.; et al. Telescope: Characterization of the Retrotranscriptome by Accurate Estimation of Transposable Element Expression. *PLoS Comput. Biol.* **2019**, *15*, e1006453. [CrossRef]

29. Grossman, R.L.; Heath, A.P.; Ferretti, V.; Varmus, H.E.; Lowy, D.R.; Kibbe, W.A.; Staudt, L.M. Toward a Shared Vision for Cancer Genomic Data. *N. Engl. J. Med.* **2016**, *375*, 1109–1112. [CrossRef]
30. Signal, B.; Kahlke, T. How_are_we_stranded_here: Quick Determination of RNA-Seq Strandedness. *BMC Bioinformatics* **2022**, *23*, 49. [CrossRef]
31. Andrews, S. *FastQC: A Quality Control Tool for High Throughput Sequence Data*; ScienceOpen, Inc.: Burlington, MA, USA, 2010.
32. Martin, M. Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads. *EMBnet J.* **2011**, *17*, 10–12. [CrossRef]
33. Langmead, B.; Salzberg, S.L. Fast Gapped-Read Alignment with Bowtie 2. *Nat. Methods* **2012**, *9*, 357–359. [CrossRef] [PubMed]
34. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R. Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/map (SAM) Format and SAMtools. *Bioinformatics* **2009**, *1000*, 2078–2079. [CrossRef] [PubMed]
35. Robinson, M.D.; McCarthy, D.J.; Smyth, G.K. edgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data. *Bioinformatics* **2010**, *26*, 139–140. [CrossRef] [PubMed]
36. Ritchie, M.E.; Phipson, B.; Wu, D.; Hu, Y.; Law, C.W.; Shi, W.; Smyth, G.K. Limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies. *Nucleic Acids Res.* **2015**, *43*, e47. [CrossRef]
37. Blighe, K.; Rana, S. Lewis EnhancedVolcano: Publication-Ready Volcano Plots with Enhanced Colouring and Labeling. *R package version* **2019**, *1*.
38. Heberle, H.; Meirelles, G.V.; da Silva, F.R.; Telles, G.P.; Minghim, R. InteractiVenn: A Web-Based Tool for the Analysis of Sets through Venn Diagrams. *BMC Bioinformatics* **2015**, *16*, 169. [CrossRef]
39. Gu, Z.; Gu, L.; Eils, R.; Schlesner, M.; Brors, B. Circlize Implements and Enhances Circular Visualization in R. *Bioinformatics* **2014**, *30*, 2811–2812. [CrossRef]
40. Chevallier, M.; Borgeaud, M.; Addeo, A.; Friedlaender, A. Oncogenic Driver Mutations in Non-Small Cell Lung Cancer: Past, Present and Future. *World J. Clin. Oncol.* **2021**, *12*, 217–237. [CrossRef]
41. Chae, Y.K.; Arya, A.; Iams, W.; Cruz, M.R.; Chandra, S.; Choi, J.; Giles, F. Current Landscape and Future of Dual Anti-CTLA4 and PD-1/PD-L1 Blockade Immunotherapy in Cancer; Lessons Learned from Clinical Trials with Melanoma and Non-Small Cell Lung Cancer (NSCLC). *J Immunother Cancer* **2018**, *6*, 39. [CrossRef]
42. Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N.S.; Wang, J.T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **2003**, *13*, 2498–2504. [CrossRef]
43. Alaimo, S.; Giugno, R.; Acunzo, M.; Veneziano, D.; Ferro, A.; Pulvirenti, A. Post-Transcriptional Knowledge in Pathway Analysis Increases the Accuracy of Phenotypes Classification. *Oncotarget* **2016**, *7*, 54572–54582. [CrossRef] [PubMed]
44. Wilkerson, M.D.; Hayes, D.N. ConsensusClusterPlus: A Class Discovery Tool with Confidence Assessments and Item Tracking. *Bioinformatics* **2010**, *26*, 1572–1573. [CrossRef]
45. Gu, Z.; Eils, R.; Schlesner, M. Complex Heatmaps Reveal Patterns and Correlations in Multidimensional Genomic Data. *Bioinformatics* **2016**, *32*, 2847–2849. [CrossRef] [PubMed]
46. Chen, F.; Zhang, Y.; Parra, E.; Rodriguez, J.; Behrens, C.; Akbani, R.; Lu, Y.; Kurie, J.M.; Gibbons, D.L.; Mills, G.B.; et al. Multiplatform-Based Molecular Subtypes of Non-Small-Cell Lung Cancer. *Oncogene* **2017**, *36*, 1384–1393. [CrossRef]
47. Zare, M.; Mostafaei, S.; Ahmadi, A.; Azimzadeh Jamalkandi, S.; Abedini, A.; Esfahani-Monfared, Z.; Dorostkar, R.; Saadati, M. Human Endogenous Retrovirus Env Genes: Potential Blood Biomarkers in Lung Cancer. *Microb. Pathog.* **2018**, *115*, 189–193. [CrossRef] [PubMed]
48. Yang, C.; Guo, X.; Li, J.; Han, J.; Jia, L.; Wen, H.-L.; Sun, C.; Wang, X.; Zhang, B.; Li, J.; et al. Significant Upregulation of HERV-K (HML-2) Transcription Levels in Human Lung Cancer and Cancer Cells. *Front. Microbiol.* **2022**, *13*, 850444. [CrossRef]
49. Arroyo, M.; Bautista, R.; Larrosa, R.; Cobo, M.Á.; Claros, M.G. Biomarker Potential of Repetitive-Element Transcriptome in Lung Cancer. *PeerJ* **2019**, *7*, e8277. [CrossRef]
50. Konen, J.M.; Rodriguez, B.L.; Fradette, J.J.; Gibson, L.; Davis, D.; Minelli, R.; Peoples, M.D.; Kovacs, J.; Carugo, A.; Bristow, C.; et al. Ntrk1 Promotes Resistance to PD-1 Checkpoint Blockade in Mesenchymal Kras/p53 Mutant Lung Cancer. *Cancers* **2019**, *11*, 462. [CrossRef]
51. Kanehisa, M.; Furumichi, M.; Tanabe, M.; Sato, Y.; Morishima, K. KEGG: New Perspectives on Genomes, Pathways, Diseases and Drugs. *Nucleic Acids Res.* **2017**, *45*, D353–D361. [CrossRef]
52. Denne, M.; Sauter, M.; Armbruester, V.; Licht, J.D.; Roemer, K.; Mueller-Lantzsch, N. Physical and Functional Interactions of Human Endogenous Retrovirus Proteins Np9 and Rec with the Promyelocytic Leukemia Zinc Finger Protein. *J. Virol.* **2007**, *81*, 5607–5616. [CrossRef] [PubMed]
53. Schmitt, K.; Heyne, K.; Roemer, K.; Meese, E.; Mayer, J. HERV-K(HML-2) Rec and np9 Transcripts Not Restricted to Disease but Present in Many Normal Human Tissues. *Mobile DNA* **2015**, 6. [CrossRef] [PubMed]
54. Grandi, N.; Tramontano, E. HERV Envelope Proteins: Physiological Role and Pathogenic Potential in Cancer and Autoimmunity. *Frontiers in Microbiology* **2018**, 9. [CrossRef] [PubMed]
55. Bonaventura, P.; Alcazer, V.; Mutez, V.; Tonon, L.; Martin, J.; Chuvin, N.; Michel, E.; Boulos, R.E.; Estornes, Y.; Valladeau-Guilemond, J.; et al. Identification of Shared Tumor Epitopes from Endogenous Retroviruses Inducing High-Avidity Cytotoxic T Cells for Cancer Immunotherapy. *Science Advances* **2022**, 8. [CrossRef] [PubMed]

56. Zhou, F.; Krishnamurthy, J.; Wei, Y.; Li, M.; Hunt, K.; Johanning, G.L.; Cooper, L.J.; Wang-Johanning, F. Chimeric Antigen Receptor T Cells Targeting HERV-K Inhibit Breast Cancer and Its Metastasis through Downregulation of Ras. *Oncoimmunology* **2015**, *4*, e1047582. [CrossRef]
57. Krishnamurthy, J.; Rabinovich, B.A.; Mi, T.; Switzer, K.C.; Olivares, S.; Maiti, S.N.; Plummer, J.B.; Singh, H.; Kumaresan, P.R.; Huls, H.M.; et al. Genetic Engineering of T Cells to Target HERV-K, an Ancient Retrovirus on MelanomaAncient Retrovirus Targeted by Engineered T Cells in Melanoma. *Clin. Cancer Res.* **2015**, *21*, 3241–3251. [CrossRef]
58. Vergara Bermejo, A.; Ragonnaud, E.; Daradoumis, J.; Holst, P. Cancer Associated Endogenous Retroviruses: Ideal Immune Targets for Adenovirus-Based Immunotherapy. *Int. J. Mol. Sci.* **2020**, *21*, 4843. [CrossRef]
59. Weyerer, V.; Strissel, P.L.; Stöhr, C.; Eckstein, M.; Wach, S.; Taubert, H.; Brandl, L.; Geppert, C.I.; Wullich, B.; Cynis, H.; et al. Endogenous Retroviral–K Envelope Is a Novel Tumor Antigen and Prognostic Indicator of Renal Cell Carcinoma. *Front. Oncol.* **2021**, *11*. [CrossRef]
60. Fehrenbacher, L.; Spira, A.; Ballinger, M.; Kowanetz, M.; Vansteenkiste, J.; Mazieres, J.; Park, K.; Smith, D.; Artal-Cortes, A.; Lewanski, C.; et al. Atezolizumab versus Docetaxel for Patients with Previously Treated Non-Small-Cell Lung Cancer (POPLAR): A Multicentre, Open-Label, Phase 2 Randomised Controlled Trial. *Lancet* **2016**, *387*, 1837–1846. [CrossRef]
61. Rittmeyer, A.; Barlesi, F.; Waterkamp, D.; Park, K.; Ciardiello, F.; von Pawel, J.; Gadgeel, S.M.; Hida, T.; Kowalski, D.M.; Dols, M.C.; et al. Atezolizumab versus Docetaxel in Patients with Previously Treated Non-Small-Cell Lung Cancer (OAK): A Phase 3, Open-Label, Multicentre Randomised Controlled Trial. *Lancet* **2017**, *389*, 255–265. [CrossRef]