






Article

Human Endogenous Retrovirus Expression Is Associated with Head and Neck Cancer and Differential Survival

Allison R. Kolbe ¹, Matthew L. Bendall ², Alexander T. Pearson ³, Doru Paul ⁴, Douglas F. Nixon ², Marcos Pérez-Losada ^{1,5,6} and Keith A. Crandall ^{1,5,*}

¹ Computational Biology Institute, Milken Institute School of Public Health, George Washington University, Washington, DC 20052, USA; akolbe@email.gwu.edu (A.R.K.); mlosada@gwu.edu (M.P.-L.)

² Division of Infectious Diseases, Department of Medicine, Weill Cornell Medicine, New York, NY 10021, USA; mlb4001@med.cornell.edu (M.L.B.); dnixon@med.cornell.edu (D.F.N.)

³ Department of Medicine, The University of Chicago Medicine, Chicago, IL 60637, USA; apearson5@medicine.bsd.uchicago.edu

⁴ Division of Hematology and Medical Oncology, Weill Cornell Medicine, New York, NY 10021, USA; dop9054@med.cornell.edu

⁵ Department of Biostatistics and Bioinformatics, Milken Institute School of Public Health, George Washington University, Washington, DC 20052, USA

⁶ CIBIO-InBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, Universidade do Porto, Campus Agrário de Vairão, 4485-661 Vairão, Portugal

* Correspondence: kcrandall@gwu.edu

Received: 30 June 2020; Accepted: 26 August 2020; Published: 28 August 2020



Abstract: Human endogenous retroviruses (HERVs) have been implicated in a variety of human diseases including cancers. However, technical challenges in analyzing HERV sequence data have limited locus-specific characterization of HERV expression. Here, we use the software Telescope (developed to identify expressed transposable elements from metatranscriptomic data) on 43 paired tumor and adjacent normal tissue samples from The Cancer Genome Atlas Program to produce the first locus-specific retrotranscriptome of head and neck cancer. Telescope identified over 3000 expressed HERVs in tumor and adjacent normal tissue, and 1078 HERVs were differentially expressed between the two tissue types. The majority of differentially expressed HERVs were expressed at a higher level in tumor tissue. Differentially expressed HERVs were enriched in members of the HERVH family. Hierarchical clustering based on HERV expression in tumor-adjacent normal tissue resulted in two distinct clusters with significantly different survival probability. Together, these results highlight the importance of future work on the role of HERVs across a range of cancers.

Keywords: RNA-seq; transposable element; endogenous retrovirus; cancer; TCGA

1. Introduction

Human endogenous retroviruses (HERVs) make up approximately 8% of the human genome, but their role in disease remains poorly understood [1–5]. Expression of human endogenous retroviruses is altered in numerous cancers, including melanoma [6,7], breast cancer [8–10], and ovarian cancer [11]. Members of multiple HERV families, including HERVH, HERVK, HERVF, HERVR, and HERVS, have been identified in cancer cell lines [12]. HERVs have been linked to oncogenesis at the DNA and protein level, but there is also evidence of beneficial HERV effects [13]. Therefore, the relationship between HERV expression and cancer is complex, and may have both positive and negative effects on cancer progression and clinical outcomes. Although many studies have characterized HERV expression

in cancer and other diseases, the high degree of repetitive and highly similar sequences in HERV elements have made locus-specific characterization of HERVs a significant challenge. Thus, characterizing the cancer retrotranscriptome has remained an elusive but important goal for cancer research.

Head and neck squamous cell carcinoma (HNSCC) affects more than 700,000 people per year worldwide, with a mortality rate that exceeds 50% [14]. HNSCC typically originates from the epithelial tissues of the oral cavity, larynx, oropharynx, or hypopharynx. HNSCC is a highly heterogeneous disease, with distinct subtypes of different etiologies and presenting with different molecular changes [15]. In particular, human papillomavirus (HPV) status, which is a risk factor for developing HNSCC, has also been shown to result in a distinct subtype of HNSCC [16]. Several existing studies have examined the transcriptome in HNSCC [17–19] and reviewed in Leemans et al. [15] using protein-coding and non-coding genomic annotations. A few studies have examined HERV family-level expression in the pan-cancer The Cancer Genome Atlas (TCGA, <https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga>) cohort [20–23]. Yet no study to date has examined the implications of locus-specific HERV expression specific to HNSCC.

Here, we apply Telescope [24], a computational software pipeline which provides accurate estimation of transposable element expression resolved to specific genomic locations, to 43 paired tumor and tumor-adjacent normal tissue RNA-seq datasets available from TCGA [25]. This work provides the first locus-specific retrotranscriptomic analysis of HNSCC.

2. Materials and Methods

2.1. TCGA Data

Paired-end RNA-sequencing data were obtained from TCGA [25]. A total of 43 head and neck cancer cases with paired tumor and healthy tissue samples were downloaded from the Genomic Data Commons (GDC) Data Portal [26]. Corresponding clinical data, including tumor stage, tumor location, vital status, and tobacco history, were downloaded using the R package *TCGAbiolinks* [27] and further supplemented by clinical data provided in [17].

2.2. Patient Demographics

The majority of the cancer cases were from the oral cavity (32/43); the remainder of the tumors were laryngeal in origin (11/43). Although HNSCC also occurs in the oropharynx, none of the available paired RNA-seq samples originated from oropharyngeal tumors. Tumor samples represented stages I–IVa (stage I: 2; stage II: 15; stage III: 8; stage IVa: 17; not reported: 1). Patients were majority male (29/43), and mostly current or former smokers. Patients without tobacco history data (8/43) were categorized as smokers, as in [17].

2.3. Sequence Data Quality Control and Trimming

Files were downloaded from GDC as BAM files (*.bam) and were converted to FASTQ (*.fastq) for quality control and trimming using the SamToFastq tool in Picard version 2.6 [28]. The downloaded files contained between 37,379,141 and 116,430,322 reads, with an average of $73,776,707 \pm 1,841,593$ (mean \pm standard error). Illumina adapters were removed and reads were trimmed for quality using Trimmomatic version 0.33 [29], removing leading and trailing bases below quality 3 and a 4-base sliding window with a quality threshold of 15. Reads with a length less than 36 bases after trimming were discarded. On average, 95% of reads remained after trimming, for a final range of 34,997,325 to 109,343,593 trimmed reads per sample.

2.4. Retrotranscriptome Quantification

Trimmed reads were aligned to the human genome (hg38) using Bowtie2 [30], as described in [24]. Briefly, the Bowtie2 alignment options were set to perform a local alignment search (*-very-sensitive-local*), allowing up to 100 alignments per read (*-k 100*), with a minimum alignment score threshold such

that fragments with ~95%+ sequence identity would be reported. Overall alignment rates ranged from 86 to 93%. Bowtie2 alignments were then provided to the Telescope v1.0.2 assign module using the HERV annotation provided by [24] and theta prior of 200,000. Final count numbers were loaded into DESeq2 [31] using the *DESeqDataSetFromMatrix* function, with tissue origin (larynx or oral cavity) and tissue type (normal or tumor) as variables in the model formula. The variance-stabilizing transformation was used for principal component and clustering analysis. Normalization and differential expression analysis were performed using the *DESeq* function in DESeq2 [31], which implements a negative binomial model and Wald test. A false-discovery rate (FDR) threshold of 0.05 was used for HERV expression. The full output of DESeq2 analysis of differential HERV expression is provided in Supplementary Table S1.

Hierarchical clustering of the top 100 differentially expressed HERVs was performed using pvcust [32]. Chi-squared tests were performed between the two tumor clusters and between the two normal clusters to determine statistical differences in gender, tissue type, smoking history, or early (stage I–II) vs. late (stage III–IVa) tumor stage. Statistical significance was determined at $p < 0.05$.

2.5. Transcriptome Quantification

Gene expression was quantified using kallisto v0.43.1 [33] using the Ensembl v96 transcriptome assembly [34]. The kallisto index was built using default settings [33]. Quantification was performed using kallisto quant with default settings. The resulting abundance tables were imported into R using tximport [35] and the transcript-to-gene mapping file provided with the Ensembl v96 index files (<https://github.com/pachterlab/kallisto-transcriptome-indices>). The model formula was specified as described above, with tissue origin (larynx or oral cavity) and tissue type (normal or tumor). On average, kallisto pseudoaligned 95% of reads. The variance-stabilizing transformation was used for principal component and clustering analysis. Normalization and differential expression analysis were performed using the *DESeq* function in DESeq2 [31], which implements a negative binomial model and Wald test. A FDR threshold of 0.01 was used for gene expression. The full output of DESeq2 analysis of host differential expression is provided in Supplementary Table S2.

2.6. Determination of HPV Status

HPV status was evaluated using PathoScope 2.0 [36] because the TCGA metadata did not include HPV status on all patients from which RNA-seq data were collected. Trimmed reads were aligned to the representative and reference viroid and virus genome databases from GenBank (<https://www.ncbi.nlm.nih.gov/genbank/>) using the PathoScope 2.0 MAP module. Reads that aligned to the human genome (hg38) were filtered from the alignment. Taxonomic assignment was performed using the PathoScope 2.0 ID module. The R package taxonomizr [37] was used to retrieve full taxonomic lineages from taxonomy identifiers, and composition and taxonomic data were imported into the R package phyloseq [38] for further manipulation. HPV status was defined by the presence of >10 reads mapped to the genus *Alphapapillomavirus*. Two patients met the criteria for being HPV positive in this study.

2.7. Gene Set Enrichment Analysis

Gene set enrichment analysis (GSEA) [39] was performed to determine enrichment of HERV families in tumor tissue. A list of HERVs ranked by differential expression was calculated by $\text{sign}(\log_{2}FC) \times -\log_{10}(\text{adjusted } p\text{-value})$. A custom gene set for HERV families was created using the HERV annotation generated in [24]. The ranked gene list and custom gene set database were provided to GSEAPreranked (GSEA v4.0.0), and analysis was performed with 1000 permutations and the “classic” enrichment statistic. A FDR cutoff of 0.25 was used according to the author’s instructions [39].

2.8. Clustering and Survival Analysis

In order to identify and characterize subtypes of HNSCC, clustering was performed on variance-stabilized HERV expression data from tumor and normal tissue. Clustering was performed separately on each tissue type using a Euclidean distance matrix and the “ward.D2” method implemented in *hclust* [40]. In each case, clusters were identified using the function *cutree*. Tumor samples were divided into two clusters. One outlier sample was removed. Survival analysis was performed on the two clusters which incorporated 38 samples. Normal samples were divided into two clusters. Survival analysis was performed using the R package *survival* v2.44 [41,42]. According to the approach implemented in [17], survival times were censored at 5 years because most cancer-related events occur before that time. For patients recorded as alive, data were censored at last follow up. Patients without death or follow-up data were excluded from the analysis, which left 39 patients in total. Survival curves were fit using the Kaplan–Meier model, stratifying by cluster membership. Significance was determined with the log-rank test at a *p*-value threshold of 0.05, but because we conduct tests independently on normal and tumor clusters, we also employ the Benjamini–Hochberg [43] multiple test correction for the two clusters at the *p*-value threshold of 0.10. Such corrections have not been routinely performed in prior work on TCGA data [17,44,45]. Resulting curves were plotted using the *ggsurvplot* function in *survminer* [46].

3. Results

3.1. HERV Expression

Using Telescope [24], we identified 3520 HERVs expressed at a minimum of 5 counts in at least 2 samples across tumor and adjacent normal tissue from HNSCC cancer cases. Tumor tissue expressed significantly more HERVs than adjacent normal tissue; on average, 1846 HERVs were expressed in tumor samples compared to 1550 HERVs expressed in normal samples. The HERV expression pattern between tumor and normal tissue is distinct, as shown by principal component analysis (Figure 1A).

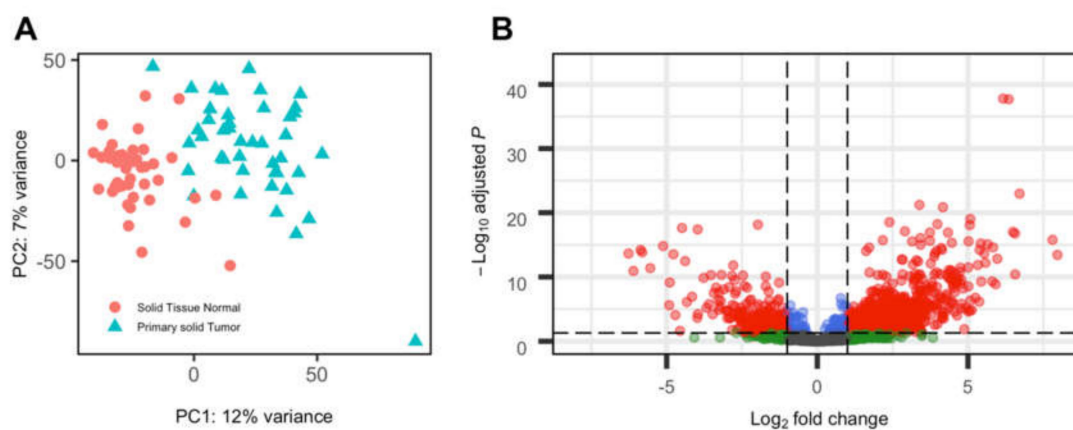


Figure 1. Human endogenous retrovirus (HERV) expression in head and neck squamous cell carcinoma (HNSCC) tumor and paired normal samples. (A) Principal component analysis, with shape and color indicating tumor vs normal tissue. (B) Volcano plot of differential expression between tumor and normal tissue. Positive log-fold change indicates higher expression in tumor tissue. The dashed horizontal line indicates adjusted *p*-value threshold of 0.05. The dashed vertical line indicates a log₂-fold change of ± 1.5 .

Out of 3520 identified HERVs, 1078 were differentially expressed between tumor and normal samples. Of these, the majority (802) were expressed at a higher level in tumor tissue compared to normal tissue (Figure 1B).

GSEA was performed to determine whether differentially expressed HERVs were enriched in particular HERV families. Among upregulated HERVs in tumor tissue, the HERVH and HARLEQUIN families were significantly enriched at $q < 0.25$ (Supplementary Table S3). The HERVE, HML6, PRIMA4, and HERVEA families were all significantly enriched ($q < 0.25$) among downregulated HERVs (Supplementary Table S4).

3.2. HERV Expression, Phenotypic and Genotypic Associations

The expression signature of top differentially expressed HERVs differed based on the tissue of origin. Clustering of the top 100 differentially expressed HERVs resulted in two distinct tumor clusters and two distinct normal clusters (Figure 2), which had significantly different tissue origins (chi-squared test, $p < 0.05$). In both cases, one cluster was dominated by oral cavity tumors (normal cluster: 10/10; tumor cluster: 22/23), which included mouth and tongue tumors, as well as tumors which overlapped the lip, oral cavity, and pharynx. The second cluster in both cases was split between oral cavity and laryngeal tumors (normal cluster: 19/30 oral cavity; tumor cluster: 13/23 oral cavity). The sublocation of oral cavity tumors varied between the two clusters; those which clustered with laryngeal tumors were more likely to be from the tongue; in contrast, the majority of oral cavity tumors in the other cluster corresponded to tumors which overlapped the lip, oral cavity, and pharynx. Smoking status, gender, age at diagnosis, and tumor stage (early vs. late) were not significantly different between the two tumor clusters or between the two normal clusters. Three normal tissue samples clustered with tumor samples. In all three cases, the paired tumor sample was found in the same cluster, and in two (HNSC06 and HNSC18), the tumor and normal sample clustered tightly together. These results may be indicative of tumor-like expression patterns in some tumor-adjacent tissues. Although HPV status is an important risk factor in HNSCC, only two patients from this cohort were HPV positive (samples HNSC16 and HNSC27). Interestingly, HPV-positive patients in this analysis did not cluster together, indicating that other variables were driving the expression patterns in these cases.

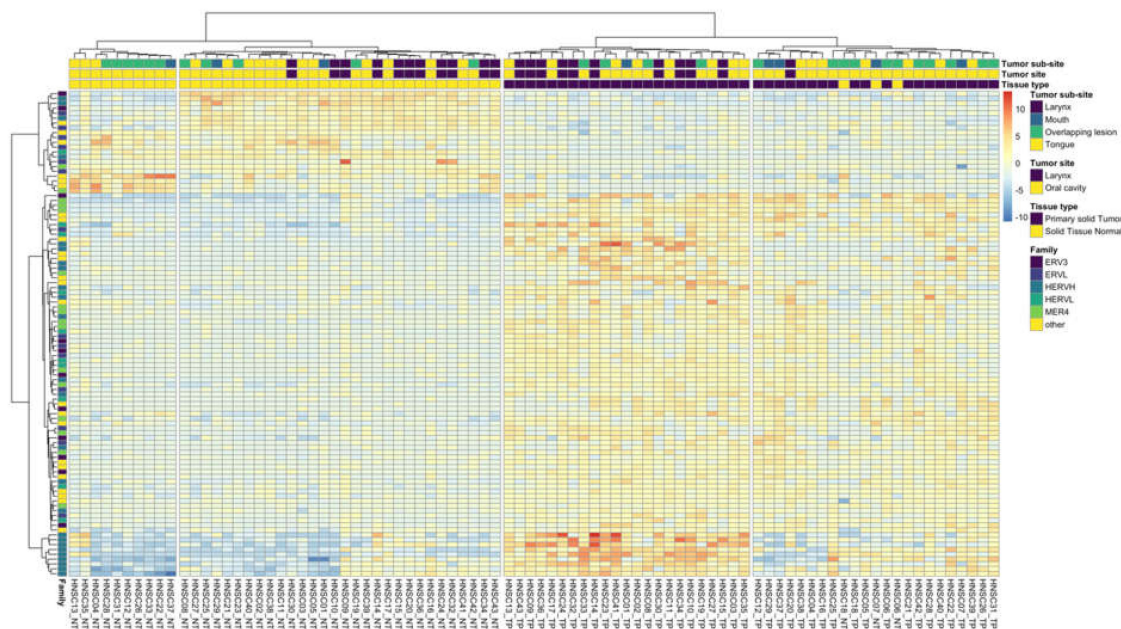


Figure 2. Expression for the top 100 differentially expressed HERVs. Heatmap colors represent the difference between each sample’s expression level and the mean expression level of each HERV (sample expression– mean expression). Row colors indicate HERV family membership. HERV families with five or fewer members present in the top 100 differentially expressed HERVs are presented as “other”.

Clustering of the top 100 differentially expressed HERVs was performed on variance-stabilized expression data using the method ward.D2 in R [40]. An ordered list of the HERV loci shown in this figure is available in Supplementary Table S5. Heatmap was generated with pheatmap [47], with the viridis [48] and RColorBrewer color palette [49].

In order to compare HERV expression with non-HERV gene expression, we identified differentially expressed host genes with kallisto [33]. Among significantly differentially expressed genes were many genes identified in previous studies, including *TP63*, *CDKN2A*, and *FADD* [17,50,51]. Similar to HERV expression, principal component analysis showed distinct expression patterns between tumor and normal tissue (Supplementary Figure S1A). As previously described [17], many genes are differentially expressed between these tissue types. Out of 33,260 genes expressed in this dataset, 4348 were significantly upregulated and 4295 were significantly downregulated (FDR < 0.01) in tumor tissue (Supplementary Figure S1B). Differentially expressed genes were found throughout the genome, similar to patterns observed for HERVs (Figure 3).

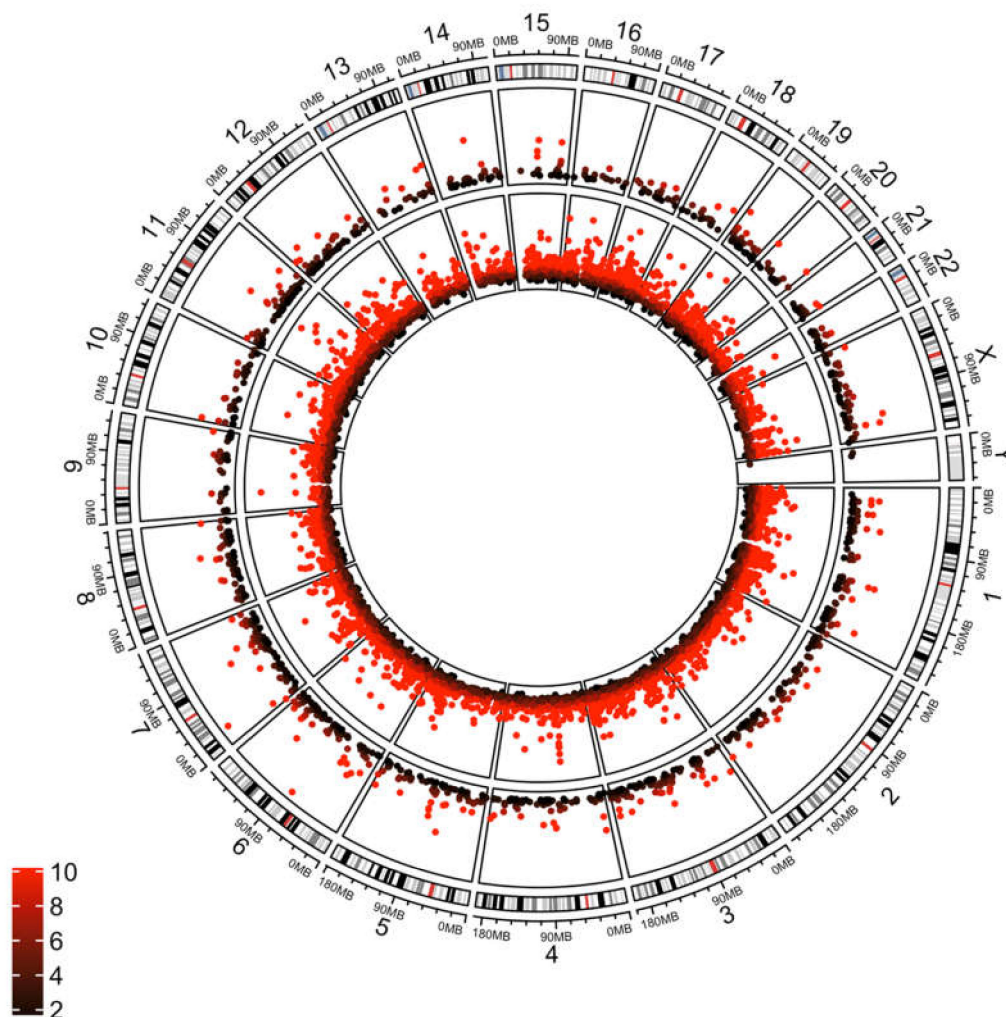


Figure 3. Circos plot showing differentially expressed host genes (inner ring) and HERVs (outer ring) across chromosomal locations. Values are plotted on a $-\log$ (adjusted p -value) scale. Color indicates degree of significance, with the darkest colors corresponding to lowest significant p -value (0.05), with the brightest red corresponding to an adjusted p -value of $p < 0.0000000001$. Plot was generated using circize in R [52].

3.3. Survival Analysis

In order to evaluate the impact of HERV expression patterns on subtypes of HNSCC which are associated with patient survival probability, we performed hierarchical clustering on the HERV expression data followed by survival analysis. For this analysis, hierarchical clustering was performed separately on tumor and normal tissue, using all HERV expression data.

Two clusters were identified from tumor tissue after removing a single outlier sample. Gender, smoking status, and tumor stage (early vs. late, as described previously) were not significantly different between these two clusters. Differences in tissue origin were significant, with one cluster composed entirely of oral cavity tumors, and the other split between laryngeal and oral cavity tumors (chi-squared test, $p < 0.05$). These clusters were also evident in the principal component analysis (Supplementary Figure S2). For the two tumor clusters, probability of patient survival was not significantly related to cluster membership (Supplementary Figure S3).

Similarly, two clusters were identified from normal tissue (Supplementary Figure S4). However, gender, tumor stage, tissue origin, and smoking status were not significantly different between these clusters (chi-squared test, $p < 0.05$). Patient survival probability differed significantly between the two normal tissue clusters (Figure 4). There was no significant enrichment of any HERV family between the two clusters, as determined by GSEA.

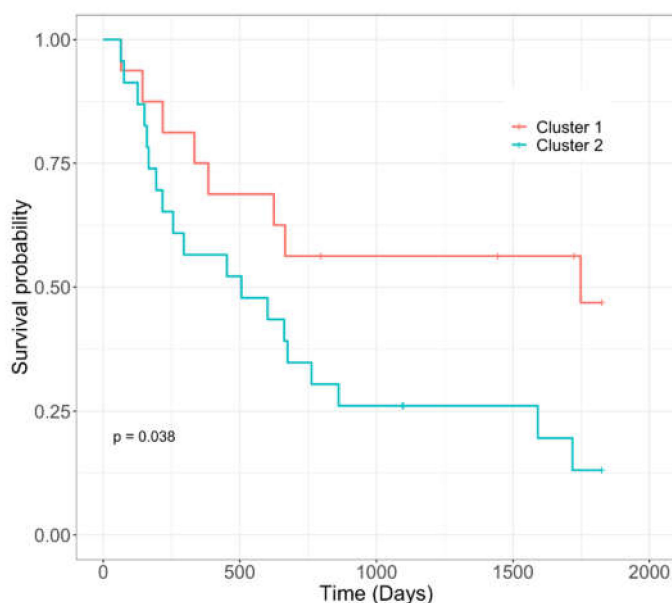


Figure 4. Survival analysis of HNSCC subtypes resulting from hierarchical clustering of HERV expression in tumor-adjacent normal tissue (Supplementary Figure S4). Vertical lines indicate censored data. Curves were fit using the Kaplan–Meier model and statistical significance was calculated using the log-rank test.

4. Discussion

HERVs have been implicated in a wide range of cancers, including breast, prostate, lymphoma, melanoma, and ovarian cancers (reviewed in [13]). Much of the work linking HERVs and cancer has focused on the HERV-K family, the most active family of HERVs, many of which are full length or nearly full length [13]. HERVK (HML2) has even been identified as a biomarker for breast cancer [53,54]. Interestingly, HERVK (HML2) was not implicated in our analysis of HNSCC cancers, whereas HERVK (HML6) was identified as being significantly enriched in the downregulated set of HERVs, indicating that although the HERVK has been implicated in many other types of cancers, not all members of this superfamily of retroelements are involved in all cancers.

In our analysis, the HERVH family was significantly enriched in tumor tissue in HNSCC. Although HERVH is not the most commonly implicated HERV in cancer, HERVH has been identified in various cancer cell lines [12,55] and pan-cancer studies [23]. Kong et al. [23] observed overexpression of HERVH-internal, the proviral portion of HERVH, in six cancer types including HNSCC. HERVH is one of the most abundant HERV families, with thousands of HERVH elements in the human genome [56]. Despite its prevalence, very few HERVH genes encode intact envelope proteins [56]. Future work should assess whether HERVH is a bystander or serves a functional role in HNSCC.

Other HERV families were found at either higher levels (HARLEQUIN) or lower levels (HERVE, HML6, PRIMA4, and HERVEA) in tumor tissues compared to adjacent normal tissues. Little is known about the role of these HERV families in HNSCC or other cancers. Future work should explore potential roles of these smaller HERV families in HNSCC.

Here, we show that HERV expression patterns as defined by hierarchical clustering in tumor-adjacent normal tissue is related to altered probability of patient survival. A similar result was previously shown by [57], who identified two clusters of non-HERV gene expression in normal tissue which were associated with differential survival. Similar to our findings, differential survival was not observed when clustering expression from tumor tissue. Although it may seem counterintuitive that expression from tumor tissue does not predict patient survival and tumor-adjacent normal tissue does, similar results have been found by a number of research groups. Transcriptional profiles were analyzed from six TCGA datasets and found that tumor-adjacent normal tissue was more predictive of patient survival than tumor tissue [58]. The authors hypothesized that tumor-adjacent normal tissue may reflect a patient's overall immunity or metabolic level, and therefore may be more informative of patient outcomes than genome dysregulation found in tumor tissue [58]. Tumor-adjacent normal tissue is known to be morphologically and phenotypically distinct from normal tissue and possesses molecular alterations that make it a unique intermediate between tumor tissue and true normal tissue [59]. Furthermore, gene expression in tumor-adjacent normal tissue has been used to predict survival in colorectal cancer patients [60], predict clinical outcome in breast cancer [61], and identify breast cancer subtypes [62]. Our findings provide additional evidence that not only are gene patterns distinct in tumor-adjacent normal tissue, but also HERV expression patterns. Therefore, future work should evaluate the links between HERV expression and other tumor characteristics which affect patient survival.

5. Conclusions

Here, we provide the first locus-specific analysis of HERV expression in head and neck cancer. We found that many HERV loci are expressed in both tumor and tumor-adjacent normal tissue, with many differentially expressed loci between the two tissue types. Interestingly, members of the HERVH family were expressed at higher levels in tumor tissue, whereas HERVK (HML2) expression, which has been associated with other types of cancer, did not vary. Furthermore, HERV expression patterns in tumor-adjacent normal tissue, as described by hierarchical clustering, were associated with differential survival. These findings emphasize the potential differences in HERV expression patterns between different cancers and emphasize the need for future work on the role of HERVs in head and neck cancer.

Supplementary Materials: The following are available online at <http://www.mdpi.com/1999-4915/12/9/956/s1>. Figure S1: Non-HERV gene expression in tumor and paired normal samples; Figure S2: Principal component analysis of HERV expression; Figure S3: Hierarchical clustering of HERV expression in tumor tissue with survival analysis of HNSCC subtypes, Figure S4: Hierarchical clustering of HERV expression in tumor-adjacent normal tissue with survival analysis; Table S1: HERV differential expression analyzed by DESeq2; Table S2: Host gene differential expression analyzed by DESeq2; Table S3: GSEA output showing HERV families enriched among upregulated HERVs in tumor tissue; Table S4: GSEA output showing HERV families enriched among downregulated HERVs in tumor tissue; Table S5: Ordering of rows in Figure 2, resulting from hierarchical clustering of top 100 differentially expressed HERVs.

Author Contributions: Conceptualization, K.A.C. and M.L.B.; methodology, A.R.K., M.L.B., and M.P.-L.; software, A.R.K. and M.L.B.; analyses, A.R.K.; resources, K.A.C. and D.F.N.; data curation, K.A.C. and A.R.K.; writing—original draft preparation, A.R.K.; writing—review and editing, all authors; visualization,

A.R.K.; supervision, M.P.-L., D.F.N., and K.A.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Institutes of Health, grant numbers R01CA206488 (D.F.N. and K.A.C.) and UL1TR000075 (K.A.C. and M.P.-L.).

Acknowledgments: We thank the GW High Performance Computing Cluster and Adam Wong for help in installing software to run our analyses.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Flockerzi, A.; Ruggieri, A.; Frank, O.; Sauter, M.; Maldener, E.; Kopper, B.; Wullich, B.; Seifarth, W.; Muller-Lantzsch, N.; Leib-Mosch, C.; et al. Expression patterns of transcribed human endogenous retrovirus HERV-K(HML-2) loci in human tissues and the need for a HERV Transcriptome Project. *BMC Genomics* **2008**, *9*, 354. [[CrossRef](#)] [[PubMed](#)]
2. Saleh, A.; Macia, A.; Muotri, A.R. Transposable Elements, Inflammation, and Neurological Disease. *Front. Neurol.* **2019**, *10*, 894. [[CrossRef](#)] [[PubMed](#)]
3. Garcia-Montojo, M.; Doucet-O'Hare, T.; Henderson, L.; Nath, A. Human endogenous retrovirus-K (HML-2): A comprehensive review. *Crit. Rev. Microbiol.* **2018**, *44*, 715–738. [[CrossRef](#)] [[PubMed](#)]
4. Meyer, T.J.; Rosenkrantz, J.L.; Carbone, L.; Chavez, S.L. Endogenous retroviruses: With us and against us. *Front. Chem.* **2017**, *5*, 23. [[CrossRef](#)] [[PubMed](#)]
5. Weiss, R.A. Human endogenous retroviruses: Friend or foe? *APMIS* **2016**, *124*, 4–10. [[CrossRef](#)]
6. Singh, S.; Kaye, S.; Gore, M.E.; McClure, M.O.; Bunker, C.B. The role of human endogenous retroviruses in melanoma. *Br. J. Dermatol.* **2009**, *161*, 1225–1231. [[CrossRef](#)]
7. Büscher, K.; Trefzer, U.; Hofmann, M.; Sterry, W.; Kurth, R.; Denner, J. Expression of human endogenous retrovirus K in melanomas and melanoma cell lines. *Cancer Res.* **2005**, *65*, 4172–4180. [[CrossRef](#)]
8. Contreras-Galindo, R.; Kaplan, M.H.; Leissner, P.; Verjat, T.; Ferlenghi, I.; Bagnoli, F.; Giusti, F.; Dosik, M.H.; Hayes, D.F.; Gitlin, S.D.; et al. Human endogenous retrovirus K (HML-2) elements in the plasma of people with lymphoma and breast cancer. *J. Virol.* **2008**, *82*, 9329–9336. [[CrossRef](#)]
9. Wang-Johanning, F.; Frost, A.R.; Jian, B.; Epp, L.; Lu, D.W.; Johanning, G.L. Quantitation of HERV-K *env* gene expression and splicing in human breast cancer. *Oncogene* **2003**, *22*, 1528–1535. [[CrossRef](#)]
10. Downey, R.F.; Sullivan, F.J.; Wang-Johanning, F.; Ambs, S.; Giles, F.J.; Glynn, S.A. Human endogenous retrovirus K and cancer: Innocent bystander or tumorigenic accomplice? *Int. J. Cancer* **2015**, *137*, 1249–1257. [[CrossRef](#)]
11. Wang-Johanning, F.; Liu, J.; Rycaj, K.; Huang, M.; Tsai, K.; Rosen, D.G.; Chen, D.-T.; Lu, D.W.; Barnhart, K.F.; Johanning, G.L. Expression of multiple human endogenous retrovirus surface envelope proteins in ovarian cancer. *Int. J. Cancer* **2007**, *120*, 81–90. [[CrossRef](#)] [[PubMed](#)]
12. Zhang, M.; Liang, J.Q.; Zheng, S. Expressional activation and functional roles of human endogenous retroviruses in cancers. *Rev. Med. Virol.* **2019**, *29*, e2025. [[CrossRef](#)] [[PubMed](#)]
13. Bannert, N.; Hofmann, H.; Block, A.; Hohn, O. HERVs New Role in Cancer: From Accused Perpetrators to Cheerful Protectors. *Front. Microbiol.* **2018**, *9*, 178. [[CrossRef](#)] [[PubMed](#)]
14. Bray, F.; Ferlay, J.; Soerjomataram, I.; Siegel, R.L.; Torre, L.A.; Jemal, A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **2018**, *68*, 394–424. [[CrossRef](#)] [[PubMed](#)]
15. Leemans, C.R.; Snijders, P.J.F.; Brakenhoff, R.H. The molecular landscape of head and neck cancer. *Nat. Rev. Cancer* **2018**, *18*, 269–282. [[CrossRef](#)]
16. Ang, K.K.; Harris, J.; Wheeler, R.; Weber, R.; Rosenthal, D.I.; Nguyen-Tân, P.F.; Westra, W.H.; Chung, C.H.; Jordan, R.C.; Lu, C.; et al. Human papillomavirus and survival of patients with oropharyngeal cancer. *N. Engl. J. Med.* **2010**, *363*, 24–35. [[CrossRef](#)]
17. Cancer Genome Atlas Network Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* **2015**, *517*, 576–582. [[CrossRef](#)]

18. Puram, S.V.; Tirosh, I.; Parikh, A.S.; Patel, A.P.; Yizhak, K.; Gillespie, S.; Rodman, C.; Luo, C.L.; Mroz, E.A.; Emerick, K.S.; et al. Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. *Cell* **2017**, *171*, 1611–1624.e24. [CrossRef]
19. Serafini, M.S.; Lopez-Perez, L.; Fico, G.; Licitra, L.; De Cecco, L.; Resteghini, C. Transcriptomics and Epigenomics in head and neck cancer: Available repositories and molecular signatures. *Cancers Head Neck* **2020**, *5*, 2. [CrossRef]
20. 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]
21. 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] [PubMed]
22. Attig, J.; Young, G.R.; Hosie, L.; Perkins, D.; Encheva-Yokoya, V.; Stoye, J.P.; Snijders, A.P.; Ternette, N.; Kassiotis, G. LTR retroelement expansion of the human cancer transcriptome and immunopeptidome revealed by de novo transcript assembly. *Genome Res.* **2019**, *29*, 1578–1590. [CrossRef] [PubMed]
23. 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] [PubMed]
24. Bendall, M.L.; de Mulder, M.; Lecanda-Sánchez, A.; Pérez-Losada, M.; Ostrowski, M.A.; Jones, R.B.; Mulder, L.C.F.; Reyes-Terán, G.; Crandall, K.A.; Ormsby, C.E.; et al. Telescope: Characterization of the retrotranscriptome by accurate estimation of transposable element expression. *PLoS Comput. Biol.* **2019**, *15*, e1006453. [CrossRef] [PubMed]
25. Tomczak, K.; Czerwińska, P.; Wiznerowicz, M. The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. *Contemp. Oncol.* **2015**, *19*, A68–A77. [CrossRef] [PubMed]
26. 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]
27. Colaprico, A.; Silva, T.C.; Olsen, C.; Garofano, L.; Cava, C.; Garolini, D.; Sabedot, T.S.; Malta, T.M.; Pagnotta, S.M.; Castiglioni, I.; et al. TCGAAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.* **2016**, *44*, e71. [CrossRef]
28. Picard Tools - By Broad Institute - GitHub Pages. Available online: <http://broadinstitute.github.io/picard/> (accessed on 28 August 2020).
29. Bolger, A.M.; Lohse, M.; Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **2014**, *30*, 2114–2120. [CrossRef]
30. Langmead, B.; Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **2012**, *9*, 357–359. [CrossRef]
31. Love, M.I.; Huber, W.; Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **2014**, *15*, 550. [CrossRef]
32. Suzuki, R.; Shimodaira, H. Pvcust: An R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* **2006**, *22*, 1540–1542. [CrossRef] [PubMed]
33. Bray, N.L.; Pimentel, H.; Melsted, P.; Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **2016**, *34*, 525–527. [CrossRef] [PubMed]
34. Zerbino, D.R.; Achuthan, P.; Akanni, W.; Amode, M.R.; Barrell, D.; Bhai, J.; Billis, K.; Cummins, C.; Gall, A.; Girón, C.G.; et al. Ensembl 2018. *Nucleic Acids Res.* **2018**, *46*, D754–D761. [CrossRef] [PubMed]
35. Love, M.I.; Soneson, C.; Robinson, M.D. Importing transcript abundance datasets with tximport. *Dim Txi. Inf. Rep. Sample1* **2017**, *1*, 5.
36. Hong, C.; Manimaran, S.; Shen, Y.; Perez-Rogers, J.F.; Byrd, A.L.; Castro-Nallar, E.; Crandall, K.A.; Johnson, W.E. PathoScope 2.0: A complete computational framework for strain identification in environmental or clinical sequencing samples. *Microbiome* **2014**, *2*, 33. [CrossRef] [PubMed]
37. Sherrill-Mix, S. GitHub - sherrillmix/taxonomizr: Parse NCBI taxonomy and accessions to find taxonomix assignments. Available online: <https://github.com/sherrillmix/taxonomizr> (accessed on 27 August 2020).
38. McMurdie, P.J.; Holmes, S. Phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE* **2013**, *8*, e61217. [CrossRef] [PubMed]

39. Subramanian, A.; Tamayo, P.; Mootha, V.K.; Mukherjee, S.; Ebert, B.L.; Gillette, M.A.; Paulovich, A.; Pomeroy, S.L.; Golub, T.R.; Lander, E.S.; et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 15545–15550. [[CrossRef](#)]
40. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2019.
41. Therneau, T.M. GitHub – therneau/survival: Survival package for R. Available online: <https://github.com/therneau/survival> (accessed on 27 August 2020).
42. Therneau, T.M.; Grambsch, P.M. *Modeling Survival Data: Extending the Cox Model*; Springer Science & Business Media: New York, NY, USA, 2013.
43. Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Series B Stat. Methodol.* **1995**, *57*, 289–300. [[CrossRef](#)]
44. Ricketts, C.J.; De Cubas, A.A.; Fan, H.; Smith, C.C.; Lang, M.; Reznik, E.; Bowlby, R.; Gibb, E.A.; Akbani, R.; Beroukhi, R.; et al. The Cancer Genome Atlas Comprehensive Molecular Characterization of Renal Cell Carcinoma. *Cell Rep.* **2018**, *23*, 3698. [[CrossRef](#)]
45. Berger, A.C.; Korkut, A.; Kanchi, R.S.; Hegde, A.M.; Lenoir, W.; Liu, W.; Liu, Y.; Fan, H.; Shen, H.; Ravikumar, V.; et al. A comprehensive pan-cancer molecular study of gynecologic and breast cancers. *Cancer Cell* **2018**, *33*, 690–705. [[CrossRef](#)]
46. Kassambara, A.; Kosinski, M.; Biecek, P. GitHub – kassambara/survminer: Survival Analysis and Visualization. Available online: <https://github.com/kassambara/survminer/> (accessed on 27 August 2020).
47. Kolde, R. pheatmap: Pretty Heatmaps. Available online: <https://cran.r-project.org/web/packages/pheatmap/index.html> (accessed on 27 August 2020).
48. Garnier, S.; Ross, N.; Rudis, B.; Sciaini, M.; Scherer, C. viridis: Default Color Maps from ‘matplotlib’. Available online: <https://cran.r-project.org/web/packages/viridis/> (accessed on 27 August 2020).
49. Neuwirth, E. RColorBrewer: ColorBrewer Palettes. Available online: <https://cran.r-project.org/web/packages/RColorBrewer/index.html> (accessed on 27 August 2020).
50. Chuang, S.C.; Agudo, A.; Ahrens, W.; Anantharaman, D.; Benhamou, S.; Boccia, S.; Chen, C.; Conway, D.I.; Fabianova, E.; Hayes, R.B.; et al. Sequence Variants and the Risk of Head and Neck Cancer: Pooled Analysis in the Inhance Consortium. *Front. Oncol.* **2011**, *1*, 13. [[CrossRef](#)] [[PubMed](#)]
51. Reddy, R.B.; Bhat, A.R.; James, B.L.; Govindan, S.V.; Mathew, R.; Ravindra, D.R.; Hedne, N.; Illiyaraja, J.; Kekatpure, V.; Khora, S.S.; et al. Meta-Analyses of Microarray Datasets Identifies ANO1 and FADD as Prognostic Markers of Head and Neck Cancer. *PLoS ONE* **2016**, *11*, e0147409. [[CrossRef](#)] [[PubMed](#)]
52. Gu, Z.; Gu, L.; Eils, R.; Schlesner, M.; Brors, B. Circlize implements and enhances circular visualization in R. *Bioinformatics* **2014**, *30*, 2811–2812. [[CrossRef](#)] [[PubMed](#)]
53. Wang-Johanning, F.; Li, M.; Esteva, F.J.; Hess, K.R.; Yin, B.; Rycaj, K.; Plummer, J.B.; Garza, J.G.; Amb, S.; Johanning, G.L. Human endogenous retrovirus type K antibodies and mRNA as serum biomarkers of early-stage breast cancer. *Int. J. Cancer* **2014**, *134*, 587–595. [[CrossRef](#)]
54. Zhao, J.; Rycaj, K.; Geng, S.; Li, M.; Plummer, J.B.; Yin, B.; Liu, H.; Xu, X.; Zhang, Y.; Yan, Y.; et al. Expression of Human Endogenous Retrovirus Type K Envelope Protein is a Novel Candidate Prognostic Marker for Human Breast Cancer. *Genes Cancer* **2011**, *2*, 914–922. [[CrossRef](#)]
55. Liang, Q.; Xu, Z.; Xu, R.; Wu, L.; Zheng, S. Expression patterns of non-coding spliced transcripts from human endogenous retrovirus HERV-H elements in colon cancer. *PLoS ONE* **2012**, *7*, e29950. [[CrossRef](#)]
56. Yi, J.-M.; Kim, H.-M.; Kim, H.-S. Human endogenous retrovirus HERV-H family in human tissues and cancer cells: Expression, identification, and phylogeny. *Cancer Lett.* **2006**, *231*, 228–239. [[CrossRef](#)]
57. Hu, S.; Yuan, H.; Li, Z.; Zhang, J.; Wu, J.; Chen, Y.; Shi, Q.; Ren, W.; Shao, N.; Ying, X. Transcriptional response profiles of paired tumor-normal samples offer novel perspectives in pan-cancer analysis. *Oncotarget* **2017**, *8*, 41334–41347. [[CrossRef](#)]
58. Huang, X.; Stern, D.F.; Zhao, H. Transcriptional profiles from paired normal samples offer complementary information on cancer patient survival—evidence from TCGA pan-cancer data. *Sci. Rep.* **2016**, *6*, 20567. [[CrossRef](#)]
59. Aran, D.; Camarda, R.; Odegaard, J.; Paik, H.; Oskotsky, B.; Krings, G.; Goga, A.; Sirota, M.; Butte, A.J. Comprehensive analysis of normal adjacent to tumor transcriptomes. *Nat. Commun.* **2017**, *8*, 1077. [[CrossRef](#)]
60. Lee, H.-Y.; Ahn, J.B.; Rha, S.Y.; Chung, H.C.; Park, K.H.; Kim, T.S.; Kim, N.K.; Shin, S.J. High KLF4 level in normal tissue predicts poor survival in colorectal cancer patients. *World J. Surg. Oncol.* **2014**, *12*, 232. [[CrossRef](#)] [[PubMed](#)]

61. Román-Pérez, E.; Casbas-Hernández, P.; Pirone, J.R.; Rein, J.; Carey, L.A.; Lubet, R.A.; Mani, S.A.; Amos, K.D.; Troester, M.A. Gene expression in extratumoral microenvironment predicts clinical outcome in breast cancer patients. *Breast Cancer Res.* **2012**, *14*, R51. [[CrossRef](#)] [[PubMed](#)]
62. Graham, K.; Ge, X.; de Las Morenas, A.; Tripathi, A.; Rosenberg, C.L. Gene expression profiles of estrogen receptor-positive and estrogen receptor-negative breast cancers are detectable in histologically normal breast epithelium. *Clin. Cancer Res.* **2011**, *17*, 236–246. [[CrossRef](#)] [[PubMed](#)]



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).