SUPPLEMENTAL NOTES

Optimization of STAR aligner settings and hERV reference sequences

To optimize for STAR aligner settings (reference masking, multimaps, and mismatches) for capture of hERV aligned reads, we generated 100 simulated samples containing a random number of paired end 50 bp reads (1-10) from each hERV, spiked with 1,000 randomly generated low-complexity reads to test for specificity. Using this simulated sample set, we first compared of alignment to the hERV reference file, either masking or not masking for low-complexity regions (9 or more repeating single nucleotides (nts), 7 or more repeating double nts, 4 or more repeating nt patterns of 3, 3 or more repeating nts patterns of 4, 2 or more repeating patterns of 5, or 2 or more repeating nt patterns of 5) of the reference sequences. We observed a significant decrease in the number of mapped low complexity reads when masking for low complexity regions within the hERV reference file (**Figure S1A**) without significant changes to sensitivity (**Figure S1B**). We next optimized for the number of multimaps. In a paired-end 50bp RNA-seq read, the maximum similarity between any two given distinguishable reads would be 49 of 50 base pair sharing, or 98% identity. Using this 98% sequence identity threshold, we looked for hERV sequences which share >98% sequence identity with any other reference hERV and observed the largest clique to be composed of 10 unique reference hERVs (**Figure S1C**). Thus, we set the maximum number of allowed multimaps to 10, which would theoretically still allow for identification among the largest clique of closely related hERV reference sequences while filtering for reads which multimap to large numbers of reference locations. To ensure this multimapping cutoff does not decrease sensitivity of alignment, we aligned our simulated dataset to the hERV reference, with multimapping limits set to 4000 (greater than the number of hERVs within the reference). Among both simulated hERV and low complexity reads, we observed that the majority of reads mapped to only one reference location, with few simulated hERV reads mapping to >2 reference locations (**Figure S1D**). Lastly, we optimized for the number of mismatches to allow during alignment through observation of mismatch distributions among simulated hERV and low complexity reads (**Figure S1E**). The vast majority of low complexity reads aligned with 8 or greater mismatches; thus, we set the mismatch cutoff to 7 or fewer. Alignment of RNA-seq reads to hERV reference was performed using STAR aligner (v2.5.3) with a masked hERV reference, multimaps ≤ 10, and mismatches ≤ 7(60).

Optimization of Alignment Strategy

The majority of viral alignment strategies rely on pre-alignment of RNA-seq data to a human genome/transcriptome reference before identification of viral sequences. This typically allows for alignment of closely related human and viral sequences to be mapped to the most accurate reference, preventing inappropriately forced alignment of closely related transcripts to viral sequences. To test the most accurate alignment strategy for identification of hERV sequences from bulk RNA-seq data, we tested 4 potential alignment workflow strategies: 1) Simultaneous alignment to both the hERV reference and human transcriptome, 2) direct alignment to the hERV proviral reference only, 3) pre-alignment to the human genome, followed by quantification of reads which fall within known hERV coordinates, and 4) pre-alignment to the human transcriptome, followed by a secondary alignment of unmapped reads to the hERV reference. Each method was tested for sensitivity through alignment of simulated hERV read data (described above) and specificity through alignment of simulated RNA-seq data derived from the GENCODE v26 database, which contains reference transcripts of annotated human genes. All runs were performed with optimized STAR aligner settings (STAR aligner v2.5.3 with a masked hERV reference, multimaps ≤ 10, and mismatches ≤ 7). Of tested options, method 1 dramatically outperformed other methods in terms of specificity **(Figure S2A)** while maintaining a high degree of sensitivity **(Figure S2B)**. Due to its high performance, we proceeded with method 1 for our workflow.

For comparison of hERV quantification using pre-alignment to human reference, RNA-seq fastq files were aligned to hg19 genome reference using STAR v2.4.2. Coordinate sorted BAM files were run through the Bedtools v2.15.0 “pairToBed” command(65), keeping only paired reads which both intersected with reference hERV genome coordinates. Remaining reads were converted to the reference hERV transcript space, with quantification performed using Salmon v0.6.0(62). Salmon quantification files were merged to generate expression matrices, which were used for all subsequent downstream comparison analyses with direct hERV alignment quantification method.

SUPPLEMENTAL NOTES

Optimization of STAR aligner settings and hERV reference sequences

To optimize for STAR aligner settings (reference masking, multimaps, and mismatches) for capture of hERV aligned reads, we generated 100 simulated samples containing a random number of paired end 50 bp reads (1-10) from each hERV, spiked with 1,000 randomly generated low-complexity reads to test for specificity. Using this simulated sample set, we first compared of alignment to the hERV reference file, either masking or not masking for low-complexity regions (9 or more repeating single nucleotides (nts), 7 or more repeating double nts, 4 or more repeating nt patterns of 3, 3 or more repeating nts patterns of 4, 2 or more repeating patterns of 5, or 2 or more repeating nt patterns of 5) of the reference sequences. We observed a significant decrease in the number of mapped low complexity reads when masking for low complexity regions within the hERV reference file (**Figure S1A**) without significant changes to sensitivity (**Figure S1B**). We next optimized for the number of multimaps. In a paired-end 50bp RNA-seq read, the maximum similarity between any two given distinguishable reads would be 49 of 50 base pair sharing, or 98% identity. Using this 98% sequence identity threshold, we looked for hERV sequences which share >98% sequence identity with any other reference hERV and observed the largest clique to be composed of 10 unique reference hERVs (**Figure S1C**). Thus, we set the maximum number of allowed multimaps to 10, which would theoretically still allow for identification among the largest clique of closely related hERV reference sequences while filtering for reads which multimap to large numbers of reference locations. To ensure this multimapping cutoff does not decrease sensitivity of alignment, we aligned our simulated dataset to the hERV reference, with multimapping limits set to 4000 (greater than the number of hERVs within the reference). Among both simulated hERV and low complexity reads, we observed that the majority of reads mapped to only one reference location, with few simulated hERV reads mapping to >2 reference locations (**Figure S1D**). Lastly, we optimized for the number of mismatches to allow during alignment through observation of mismatch distributions among simulated hERV and low complexity reads (**Figure S1E**). The vast majority of low complexity reads aligned with 8 or greater mismatches; thus, we set the mismatch cutoff to 7 or fewer. Alignment of RNA-seq reads to hERV reference was performed using STAR aligner (v2.5.3) with a masked hERV reference, multimaps ≤ 10, and mismatches ≤ 7(60).

Optimization of Alignment Strategy

The majority of viral alignment strategies rely on pre-alignment of RNA-seq data to a human genome/transcriptome reference before identification of viral sequences. This typically allows for alignment of closely related human and viral sequences to be mapped to the most accurate reference, preventing inappropriately forced alignment of closely related transcripts to viral sequences. To test the most accurate alignment strategy for identification of hERV sequences from bulk RNA-seq data, we tested 4 potential alignment workflow strategies: 1) Simultaneous alignment to both the hERV reference and human transcriptome, 2) direct alignment to the hERV proviral reference only, 3) pre-alignment to the human genome, followed by quantification of reads which fall within known hERV coordinates, and 4) pre-alignment to the human transcriptome, followed by a secondary alignment of unmapped reads to the hERV reference. Each method was tested for sensitivity through alignment of simulated hERV read data (described above) and specificity through alignment of simulated RNA-seq data derived from the GENCODE v26 database, which contains reference transcripts of annotated human genes. All runs were performed with optimized STAR aligner settings (STAR aligner v2.5.3 with a masked hERV reference, multimaps ≤ 10, and mismatches ≤ 7). Of tested options, method 1 dramatically outperformed other methods in terms of specificity **(Figure S2A)** while maintaining a high degree of sensitivity **(Figure S2B)**. Due to its high performance, we proceeded with method 1 for our workflow.

For comparison of hERV quantification using pre-alignment to human reference, RNA-seq fastq files were aligned to hg19 genome reference using STAR v2.4.2. Coordinate sorted BAM files were run through the Bedtools v2.15.0 “pairToBed” command(65), keeping only paired reads which both intersected with reference hERV genome coordinates. Remaining reads were converted to the reference hERV transcript space, with quantification performed using Salmon v0.6.0(62). Salmon quantification files were merged to generate expression matrices, which were used for all subsequent downstream comparison analyses with direct hERV alignment quantification method.

SUPPLEMENTAL FIGURES & TABLES

Superfamily	Traditional classification
MER50LIKE	Gammaretrovirus-like (Class I)
MLLV*	Gammaretrovirus-like (Class I)
HEPSI	Gammaretrovirus-like (Class I)
HERVERI	Gammaretrovirus-like (Class I)
HERVIPADP	Gammaretrovirus-like (Class I)
HUERSP	Gammaretrovirus-like (Class I)
HERVFRDLIKE	Gammaretrovirus-like (Class I)
HML	Betaretrovirus-like (Class II)
HSERVIII	Class III
HERVW9	Gammaretrovirus-like (Class I)
HERVHF	Gammaretrovirus-like (Class I)

Table S1: Conversion table between hERV superfamily and traditional classification.

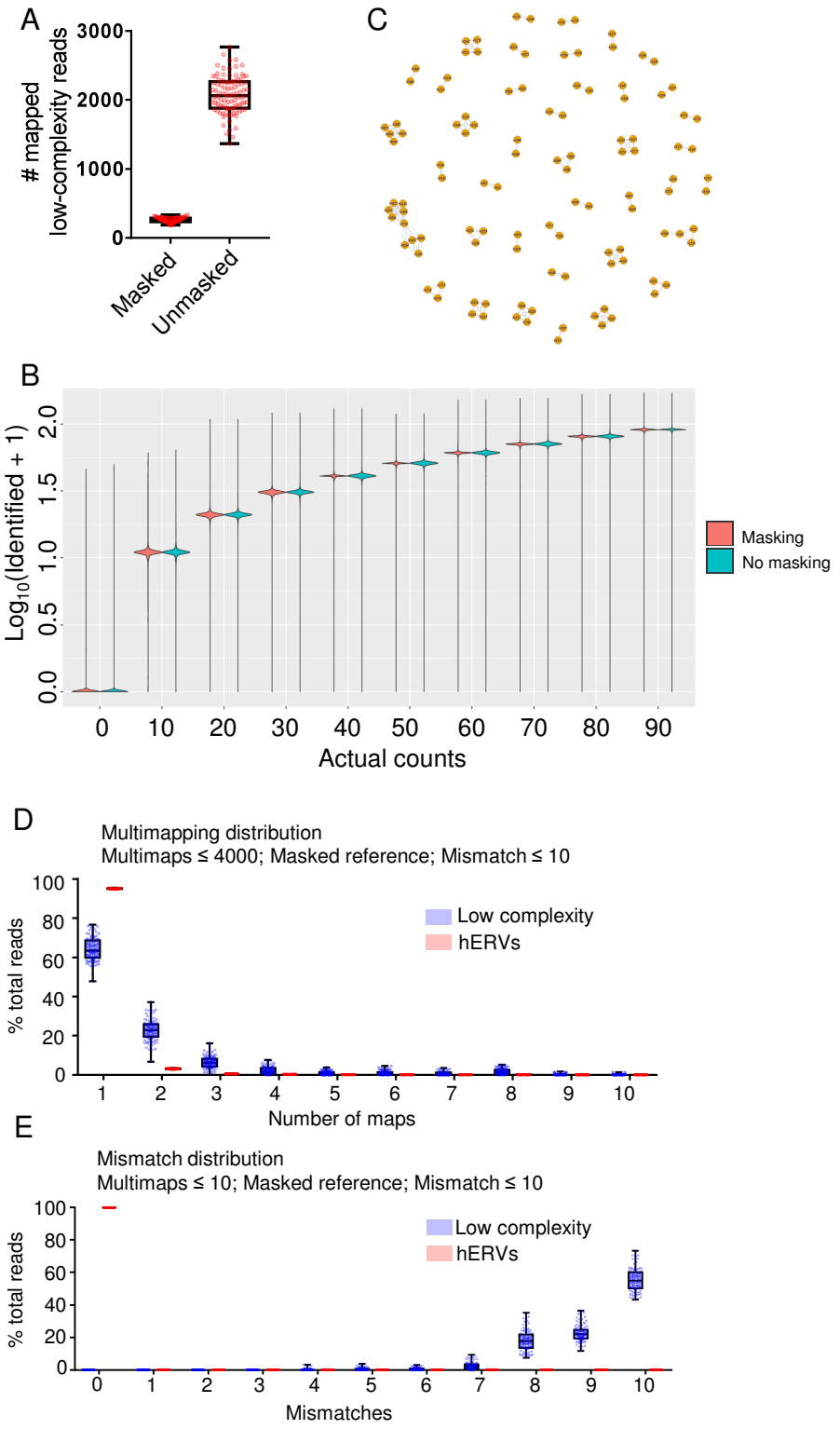


Figure S1: Optimization of STAR aligner settings and hERV reference sequences. (A) Number of mapped low complexity reads using hERV reference with masked or unmasked low complexity region. (B) Effects of hERV reference masking on accuracy of hERV quantification, displaying actual read counts within simulated hERV data (x-axis) and log10(identified reads) through alignment of simulated data (y-axis). Graph represents distribution of identified reads, with lines encompassing minimum to maximum values. (A&B) Subfigures a&b were run with STAR parameters of multimaps ≤ 10 and mismatches ≤ 7. (C) Network depiction of hERV reference sequences, where two connected hERV nodes represent >95% sequence identity using pairwise alignment of all hERV reference sequences. (D&E) Distribution of multimaps (D) and mismatches (E) through STAR alignment of simulated hERV reads with spiked-in low complexity reads. Low complexity reads (blue) and simulated hERV reads (red) are independently displayed. STAR parameters for subfigures D&E are displayed within the figure text. (A,D,& E) Data represent values (dots), median (middle line), with box encompassing the 25th to 75th percentile, and whiskers encompassing minimum to maximum values.

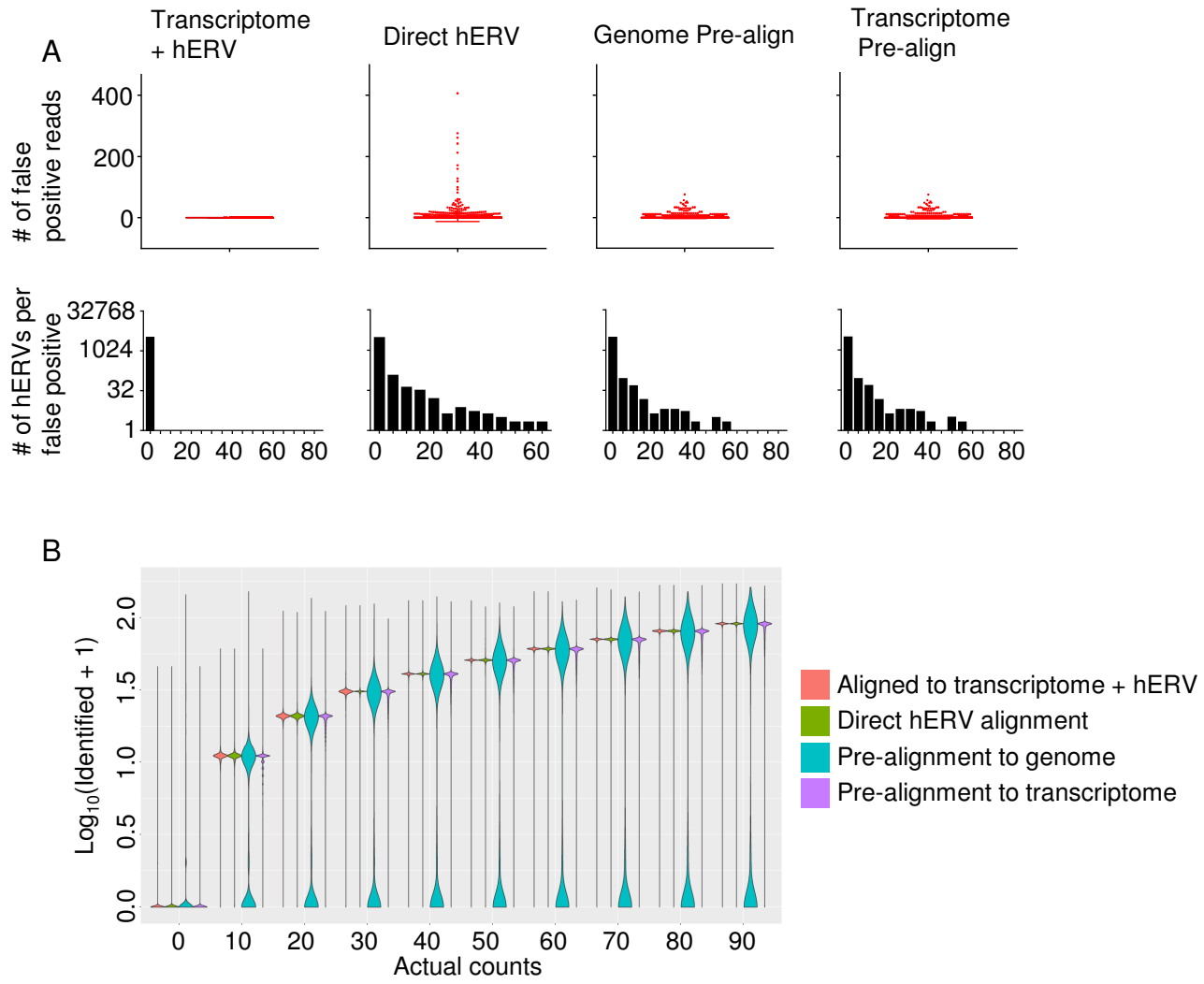


Figure S2: **Comparison of hERV alignment strategies.** (A) Number of false-positive reads identified per hERV (top) and histogram of hERV numbers per false positive reads (bottom) identified by each of the above methods. Methods include 1) Direct alignment to the hERV proviral reference only, 2) pre-alignment to the human genome, followed by quantification of reads which fall within known hERV coordinates, 3) pre-alignment to the human transcriptome, followed by a secondary alignment of unmapped reads to the hERV reference, and 4) simultaneous alignment to both the hERV reference and human transcriptome. Reads were generated using simulated RNA-seq data from GENCODE v26 sequences. (B) Quantification of hERV expression in simulated hERV reads using each of the above four methods, demonstrating the number of reads identified versus the actual number of reads contained per sample. Plot displays the actual read counts within simulated hERV data (x-axis) along with the log10(identified reads) by alignment of simulated data (y-axis). Graph represents distribution of identified reads, with lines encompassing minimum to maximum values.

Full name	Abbreviation	Cohort size
Adrenocortical carcinoma	ACC	92
Bladder urothelial carcinoma	BLCA	412
Breast invasive carcinoma	BRCA	1097
Cervical and endocervical cancers	CESC	307
Cholangiocarcinoma	CHOL	45
Colon adenocarcinoma	COAD	458
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	DLBC	48
Head and Neck squamous cell carcinoma	HNSC	528
Kidney Chromophobe	KICH	113
Kidney renal clear cell carcinoma	KIRC	537
Kidney renal papillary cell carcinoma	KIRP	291
Brain Lower Grade Glioma	LGG	515
Liver hepatocellular carcinoma	LIHC	377
Lung adenocarcinoma	LUAD	522
Lung squamous cell carcinoma	LUSC	504
Mesothelioma	MESO	87
Pancreatic adenocarcinoma	PAAD	185
Pheochromocytoma and Paraganglioma	PCPG	179
Prostate adenocarcinoma	PRAD	499
Rectum adenocarcinoma	READ	171
Sarcoma	SARC	261
Skin Cutaneous Melanoma	SKCM	470
Stomach adenocarcinoma	STAD	443
Testicular Germ Cell Tumors	TGCT	134
Thyroid carcinoma	THCA	503
Thymoma	THYM	124
Uterine Corpus Endometrial Carcinoma	UCEC	548
Uterine Carcinosarcoma	UCS	57
Uveal Melanoma	UVM	80

Table S2: Abbreviations and cohort sizes for TCGA cancer types.

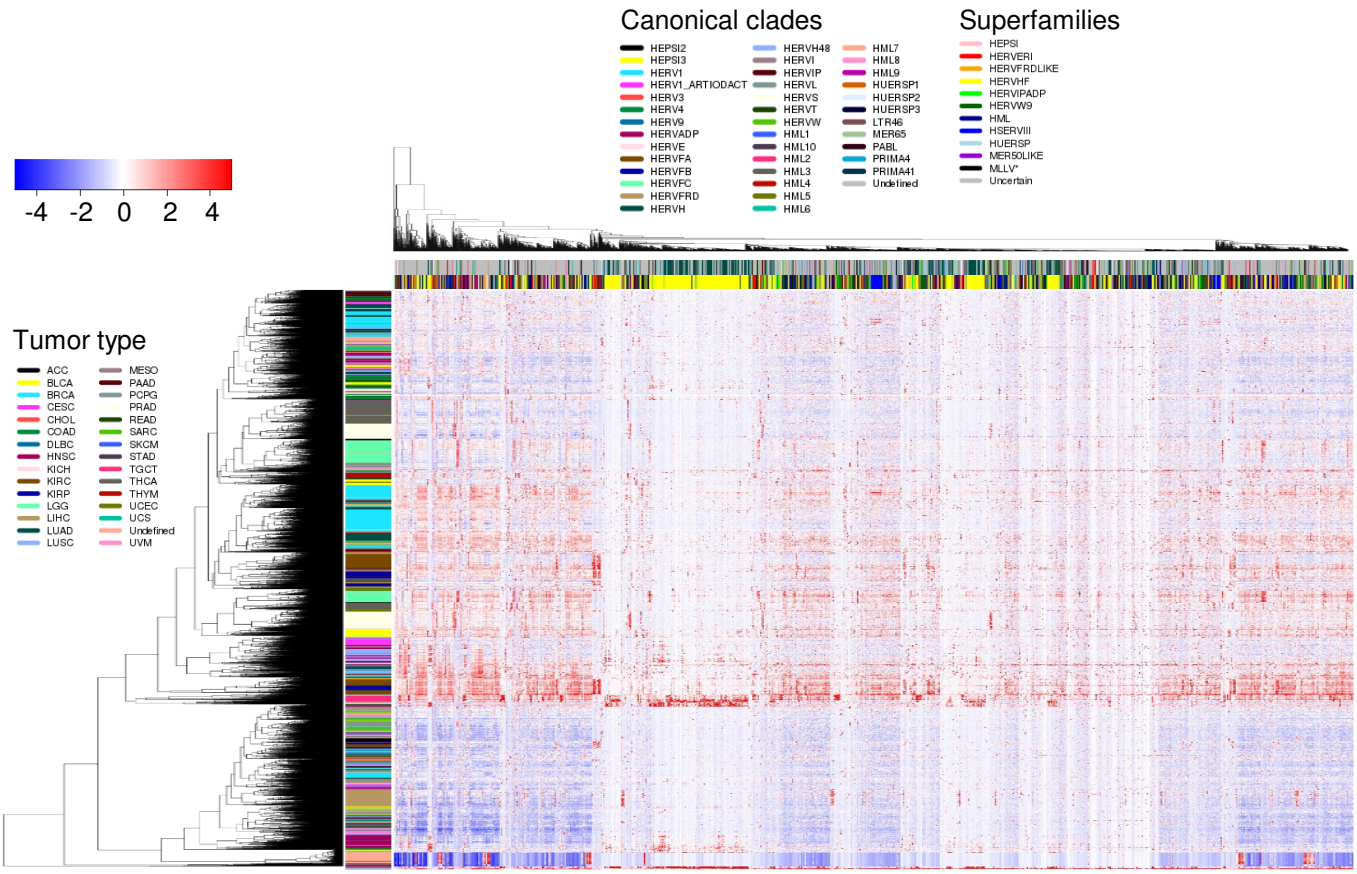


Figure S3: **Expression of hERV among TCGA pan-cancer RNA-seq dataset.** Column-side color bar displays superfamily and canonical clade classification. Row-side color bar displays TCGA tumor type. Colors represent z-score of counts, normalized by each hERV across all tumors.

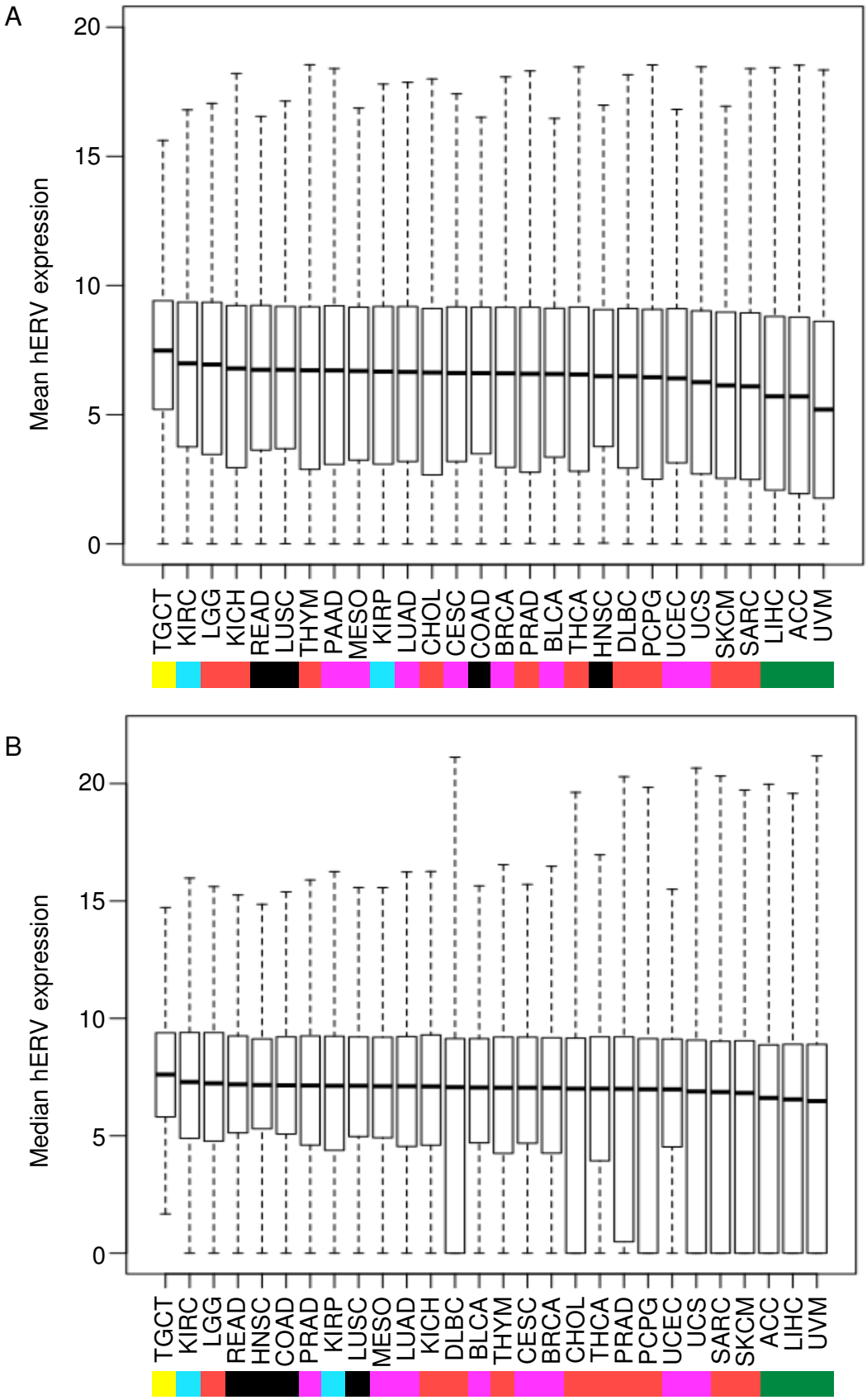


Figure S4: Mean (A) and median (B) hERV expression among the TCGA pan-cancer dataset. Data are split by tumor type, with x-axis color bars match the grouping identified by co-clustering in Figures 1B & S5, corresponding to tumor types with similar hERV expression patterns. Data represent median (middle line), with box encompassing the 25th to 75th percentile, and whiskers encompassing 1.5x the interquartile range from the box.

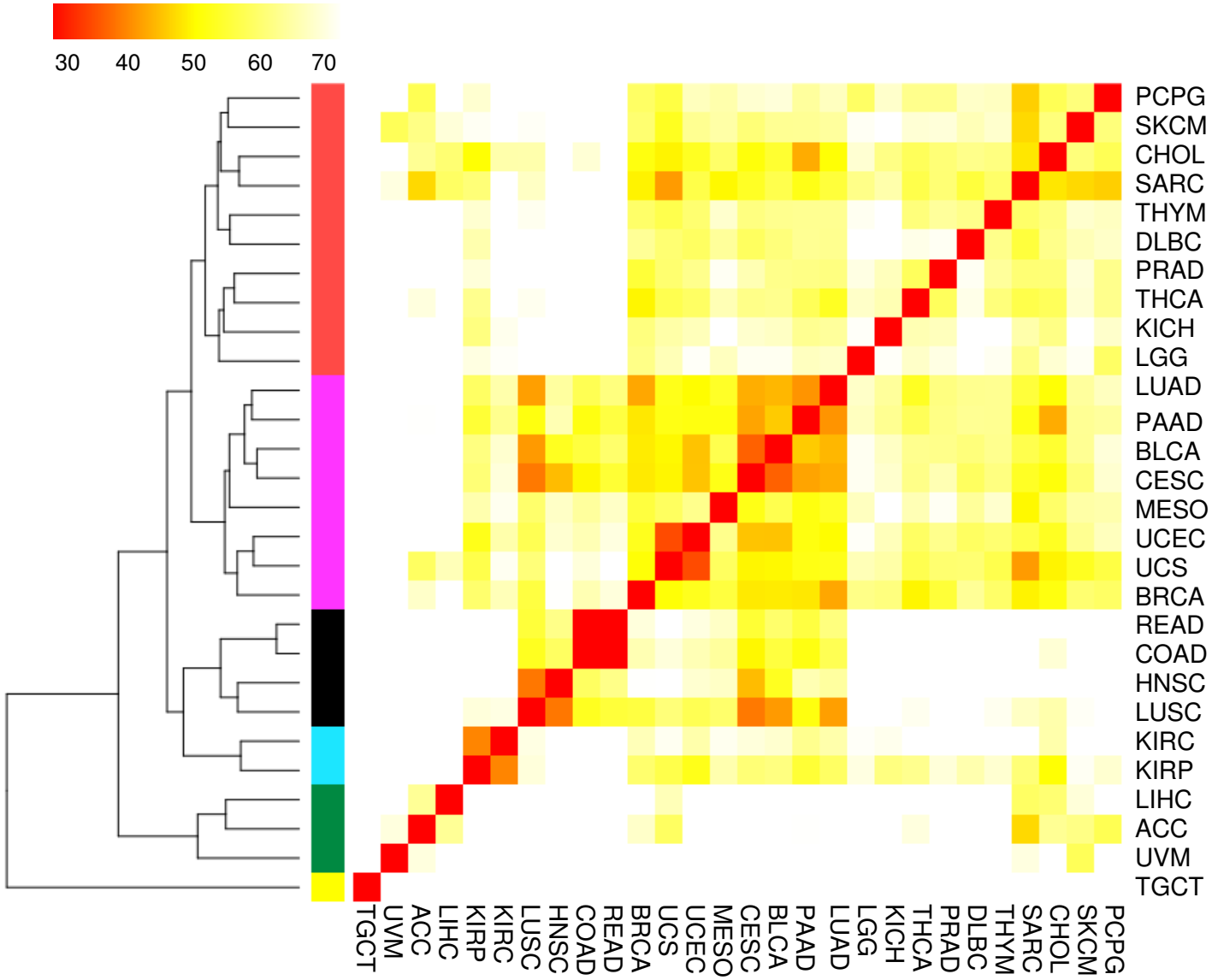


Figure S5: Unsupervised clustering of Euclidean distances in hERV expression between each pairwise TCGA cancer type. Row-side color bar represents clusters determined from a cut-tree (height = 140) of hierarchical clustering of Euclidean distance of mean hERV expression between each cancer type.

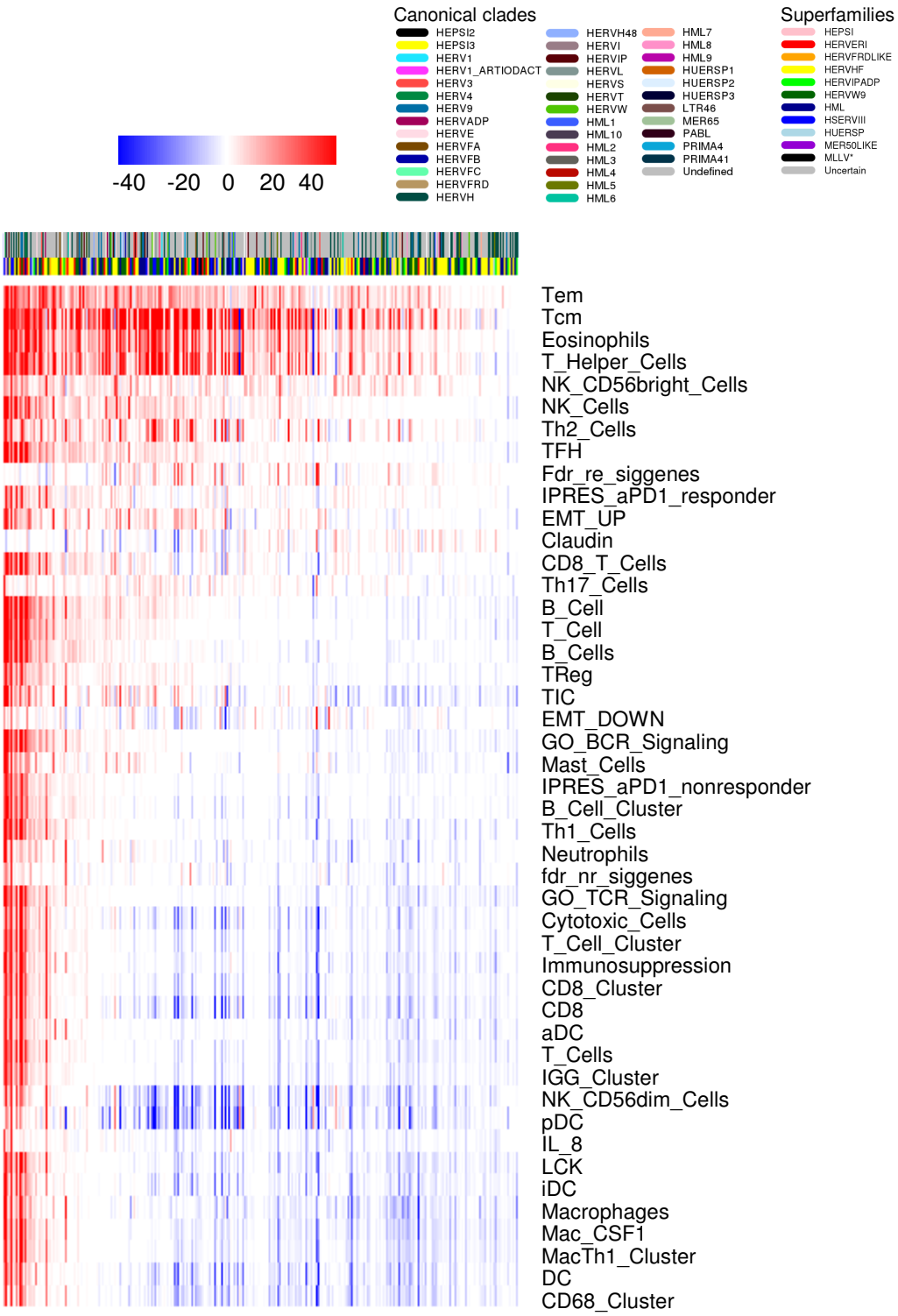


Figure S6: Heatmap of association between UQN hERV expression and immune gene signature expression among TCGA pan-cancer dataset. FDR corrected p-values (GLM) represented by intensity of color and direction of coefficient represented by color (red: positive, blue: negative). Column-side color bar displays hERV superfamily and canonical clade classifications. Rows and columns are ordered by number of significantly positive associations.

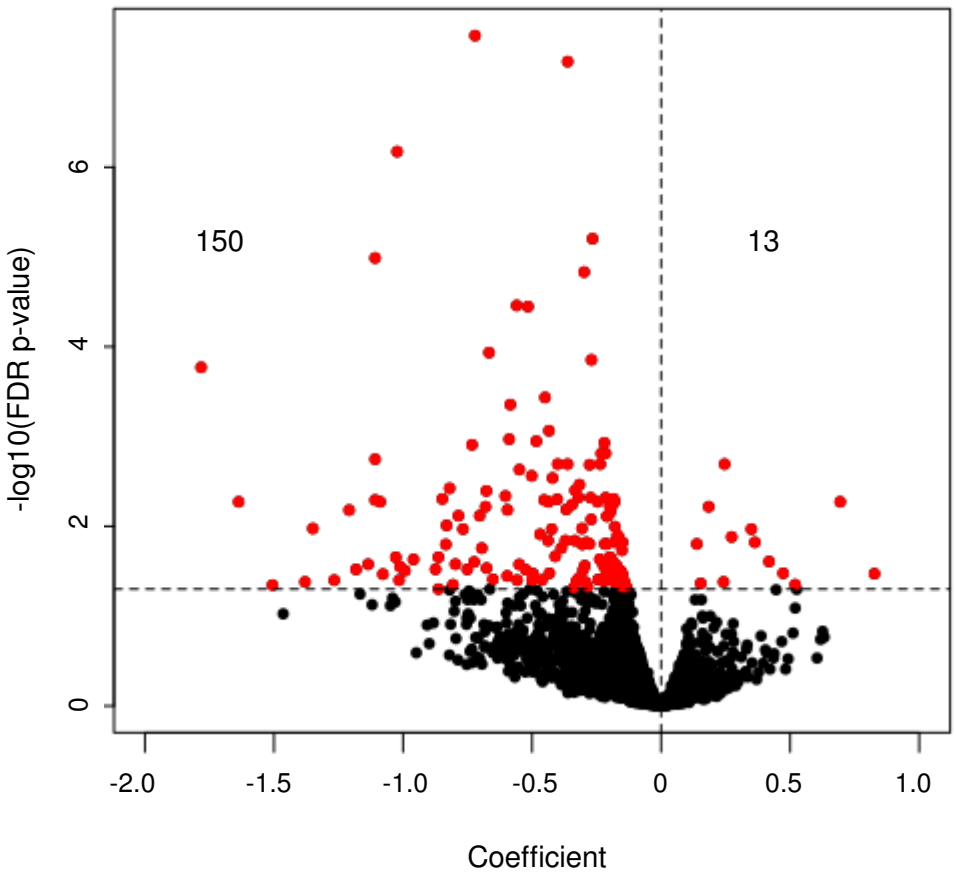


Figure S7: Association between UQN hERV expression and age TCGA pan-cancer dataset. –log10 FDR adjusted p-value (GLM) is shown along the y axis, with coefficient along the x-axis. Red dots above the dashed line represents FDR-corrected p-value < 0.05, with labels quantifying number of significant hERVs with coefficient >/< 0.

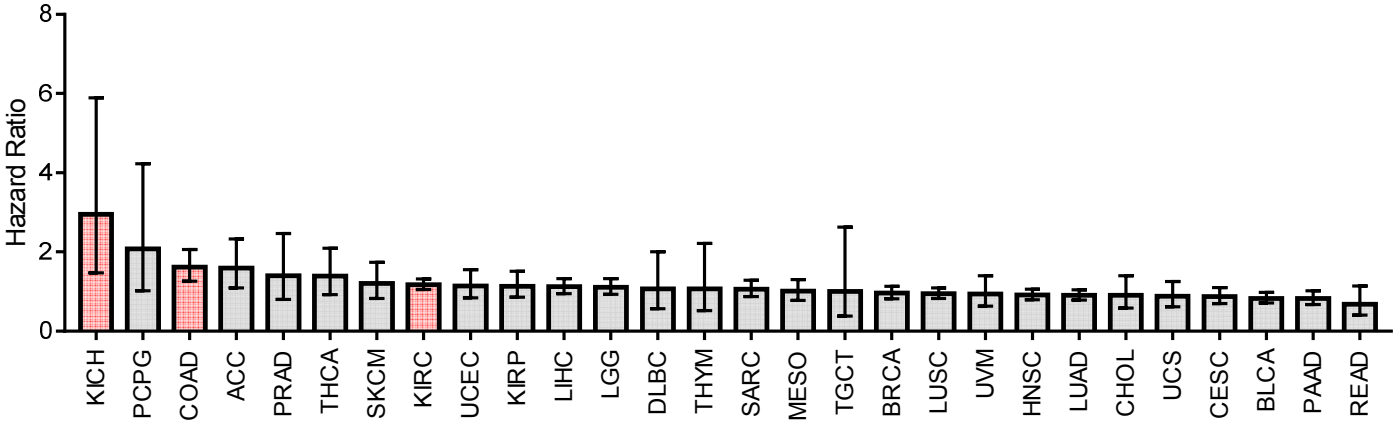


Figure S8: Boxplots of CoxPH hazard ratio for mean hERV expression as a predictor for overall survival by each tumor type. Bars colored in red have FDR-adjusted p-values < 0.05. Data represent mean +/- standard deviation for hazard ratio of each independent hERV.

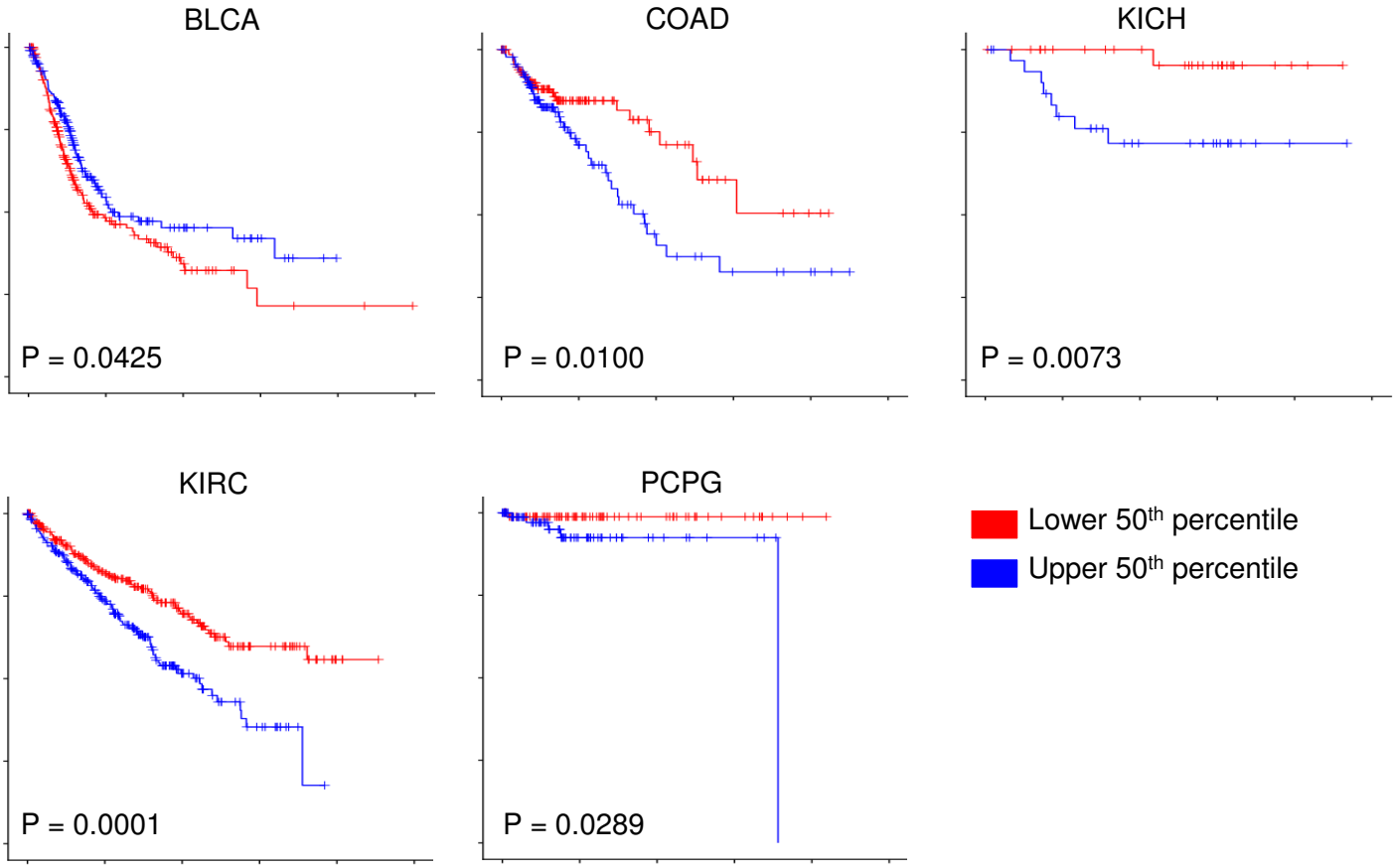


Figure S9: Kaplan-Meier curves with log-ranked p-values displayed for overall survival differences between patients within upper versus lower 50th percentile average hERV expression in TCGA BLCA, COAD, KICH, KIRC, and PCPG. Only cancer types with significant differences by log-rank analysis of overall survival are displayed.

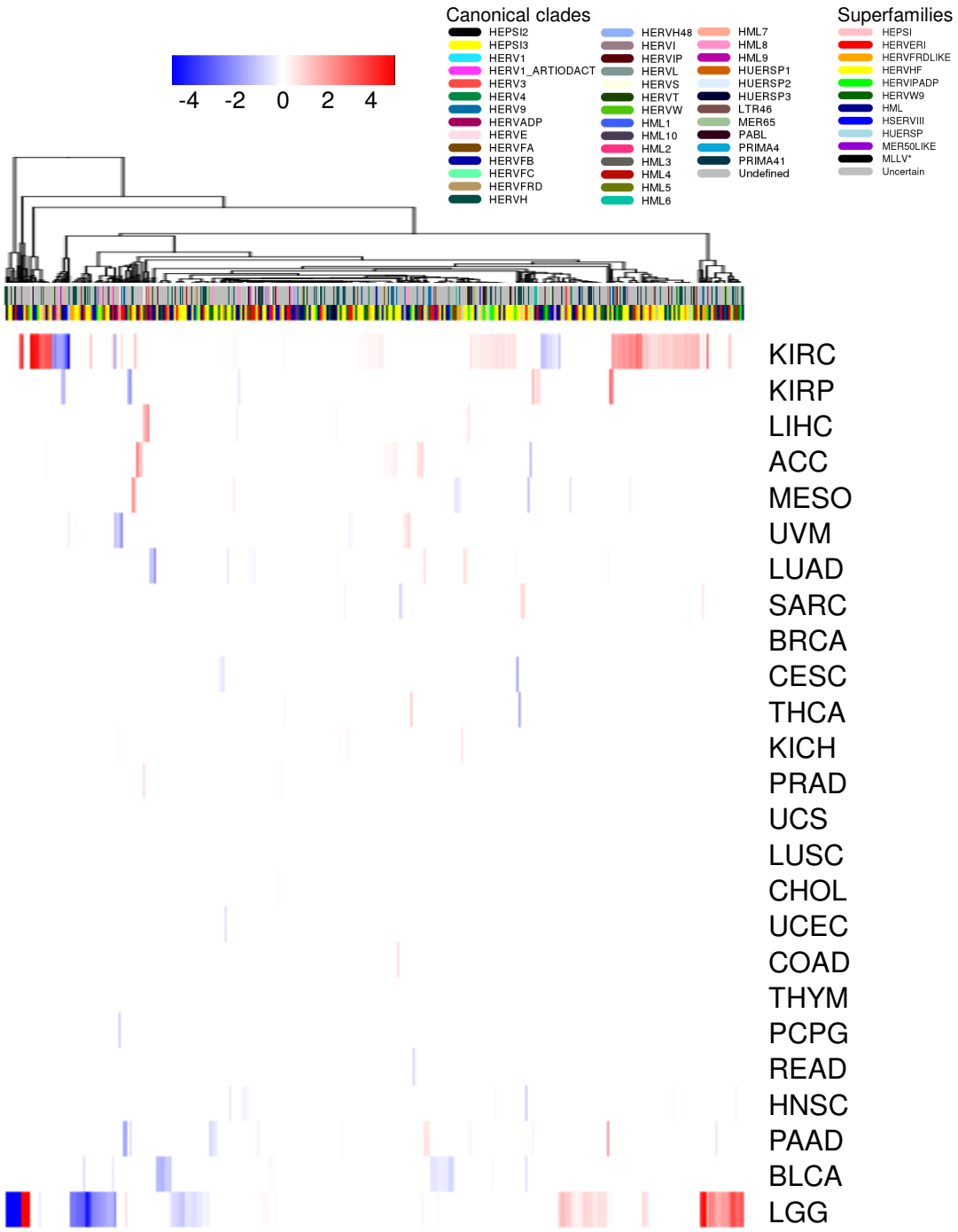


Figure S10: Heatmap of association between UQN hERV expression and survival among TCGA pan-cancer dataset. Bonferroni corrected p-values (GLM) represented by intensity of color and direction of coefficient represented by color (red: positive, blue: negative). Column-side color bar displays hERV superfamily and canonical clade classifications. Survival analysis filtered by hERVs and tumor types with at least 1 significant comparison.

Table S3
Too large to display

Table S3: Superfamily and hERV group signature classifications for all reference hERVs.

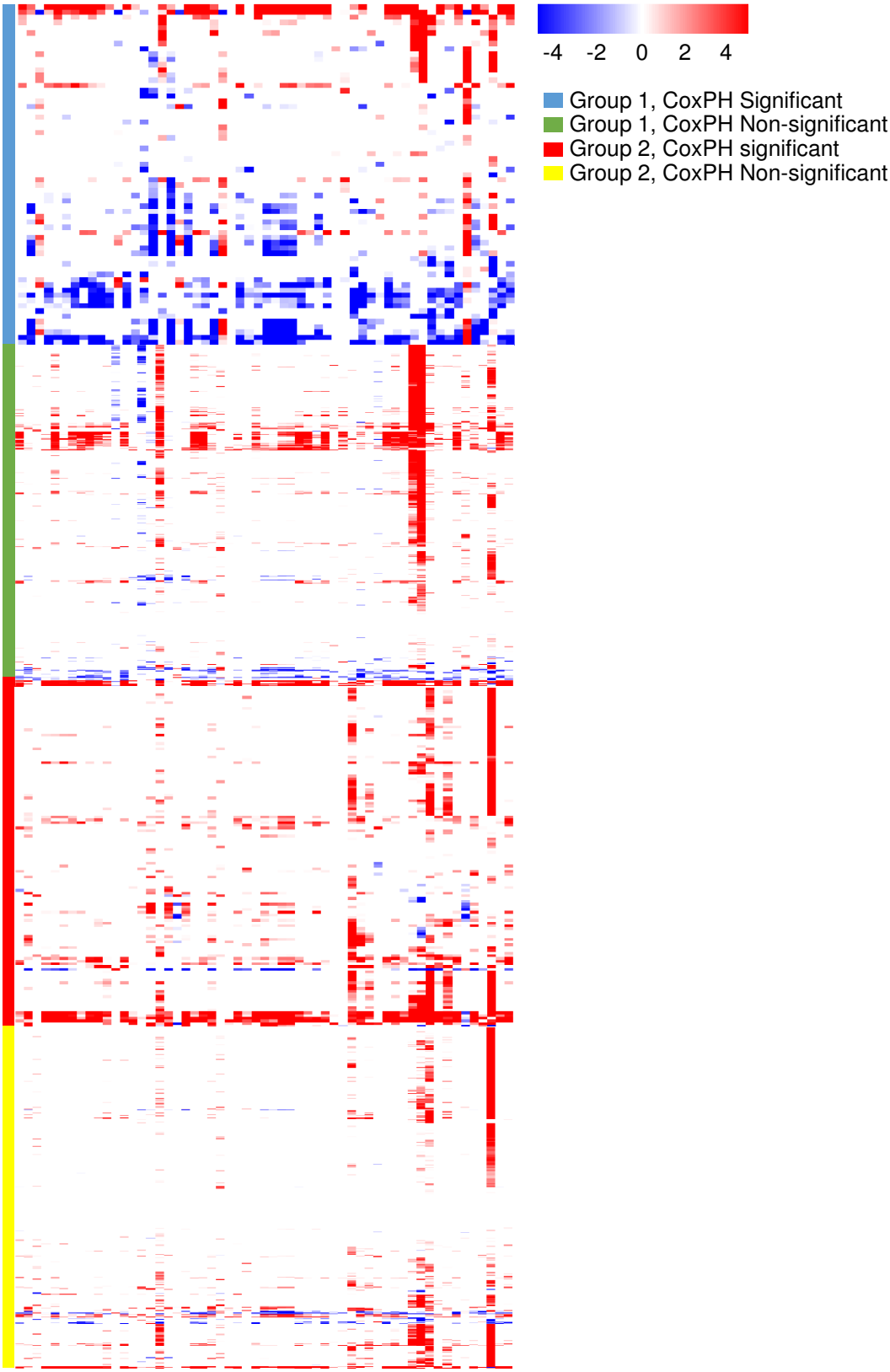


Figure S11: Heatmap of association between UQN hERV expression among CoxPH significant and non-significant, group 1 and 2 hERVs with immune gene signature expression. FDR corrected p-values represented by intensity of color and direction of coefficient represented by color (red: positive, blue: negative). hERVs (rows) are ordered by unsupervised clustering within each group, with immune gene signatures (column) unordered.

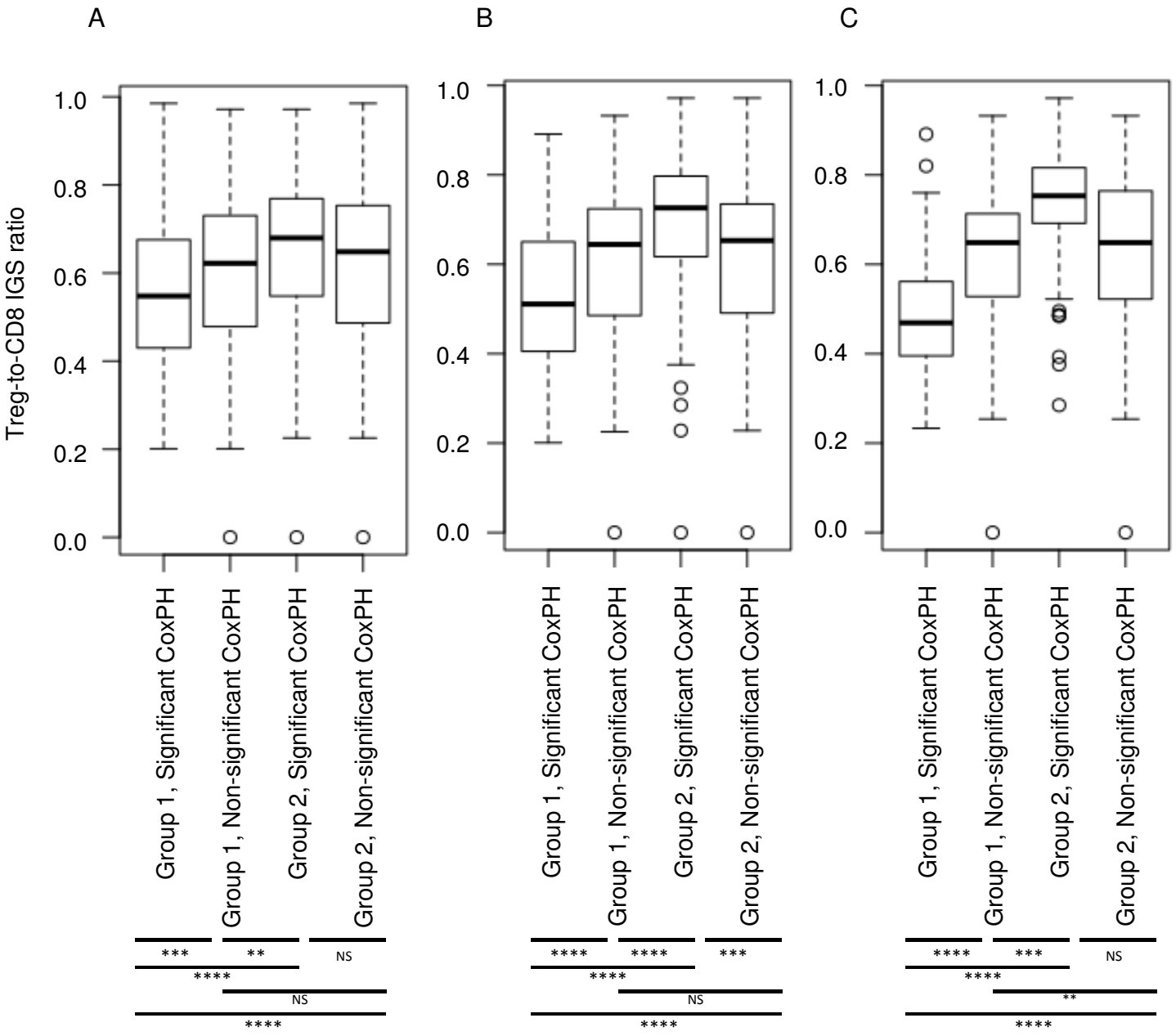
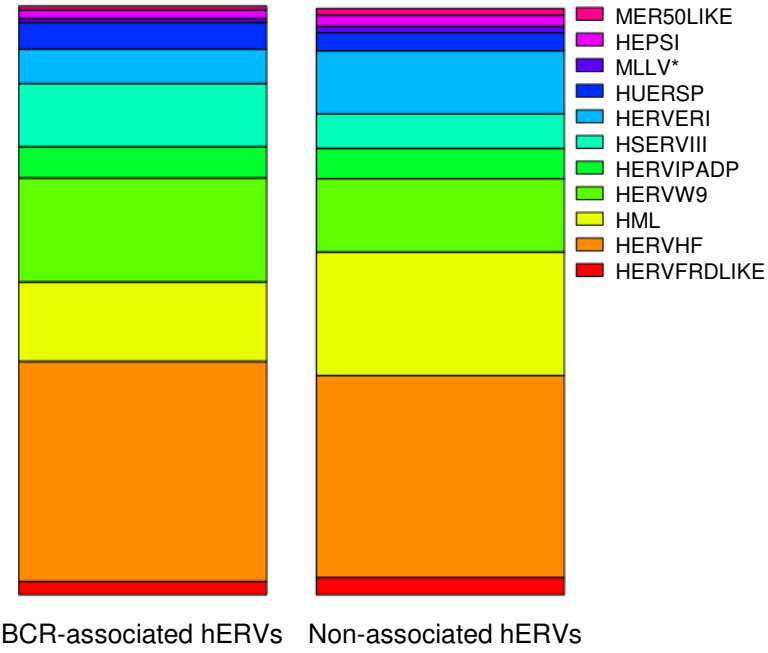


Figure S12: Treg-to-CD8+ IGS ratio expression within TCGA KIRC samples. Data split by top (A) 50th, (B) 25th, and (C) 10th percentile expression of CoxPH significant and non-significant, group 1 and 2 hERVs. Statistical analysis performed by Mann-Whitney u test (*: $p \leq 0.05$, **: $p \leq 0.01$, ***: $p \leq 0.001$, ****: $p \leq 0.0001$). Data represent median (middle line), with box encompassing the 25th to 75th percentile, whiskers encompassing 1.5x the interquartile range from the box, and outliers shown by dots.

A



B

Superfamily	Chi-square test p value
MER50LIKE	0.37
HEPSI	0.43
MLLV*	0.43
HUERSP	0.09
HERVERI	1.33×10^{-5}
HSERVIII	1.76×10^{-6}
HERVIPADP	0.94
HERVW9	1.12×10^{-4}
HML	9.10×10^{-7}
HERVHF	0.172
HERVFRDLIKE	0.39

Figure S13: (A) hERV superfamilies among BCR-associated and non-associated hERVs, with (B) FDR-corrected Chi-square test determined p-values with highlighted significant values. BCR-associated hERVs defined by hERVs significantly associated with top four B cell receptor clones displayed in figure 3A.

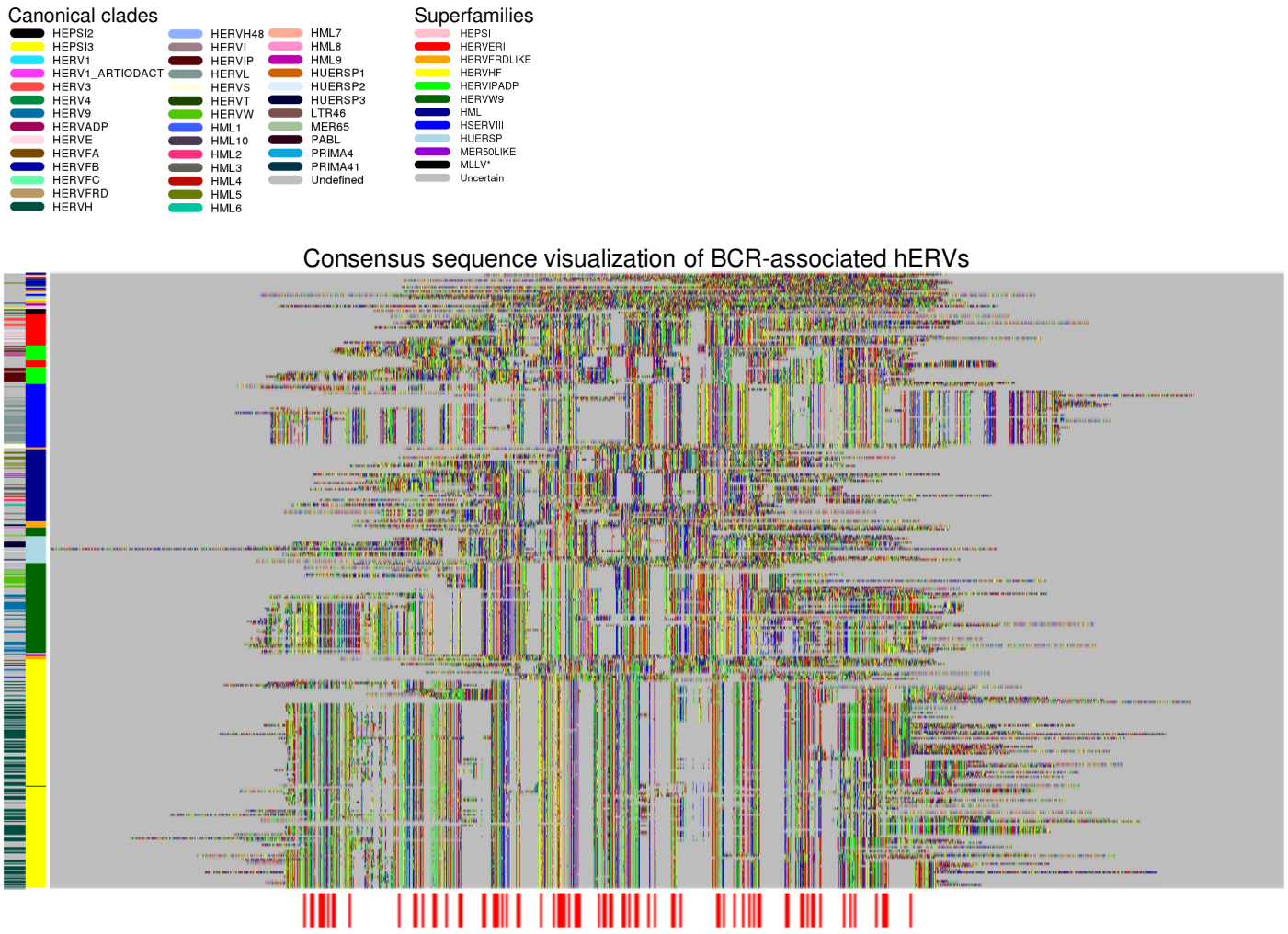


Figure S14: Visual representation of hERV provirus sequences among all significantly associated hERVs to the top four T cell clones. hERV sequences aligned by multiple sequence alignment, with base pair sequences displayed by color (A: blue; T: red; C: green; G: yellow; gap: grey). Row-side color bar represents hERV superfamily and canonical clade classifications. Column-side color bar shows areas with >25% sequence homology (red).

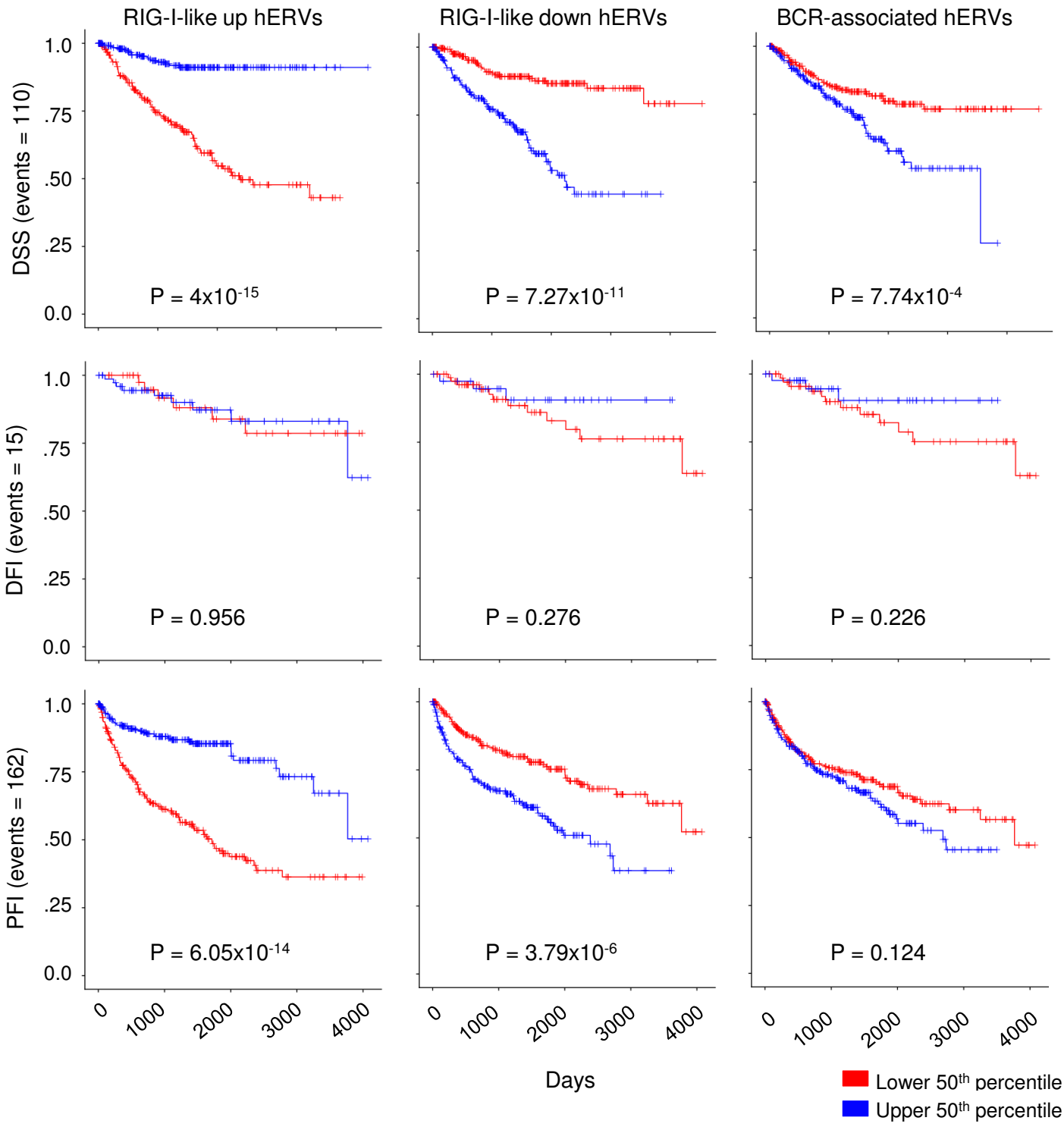


Figure S15: Kaplan-Meier survival curves for TCGA KIRC patients. Curves are defined by the upper (blue) and lower (red) 50th percentile of expression for each of the three hERV group signatures represented in Figure 5a, with curves representing disease-specific survival (DSS), disease-free interval (DFI) and progression-free interval (PFI). Proportion of total events for each survival metric displayed along y-axis. Note that the number of events for DFI is reported to be underpowered for analysis. P-values represent log-rank significance testing.

	Model features	Hazard Ratio	p-value
hERV signatures	RIG-I-like up	0.69478	0.0209
	RIG-I-like down	3.03480	$4.6x10^{-10}$
	BCR-associated	0.29343	0.0330
Clinical Stage	Stage 2	1.03692	0.9091
	Stage 3	1.77581	0.0087
	Stage 4	4.25697	$2.0x10^{-12}$
Molecular Subtype	M1	1.01424	0.9631
	M2	1.30198	0.3171
	M3	1.00516	0.9841
	M4	0.87005	0.6113

Table S4: Multivariable Cox proportional hazard ratio and significance. Results are derived from a full model composed of hERV signatures, clinical stage, and ccRCC molecular subtype.

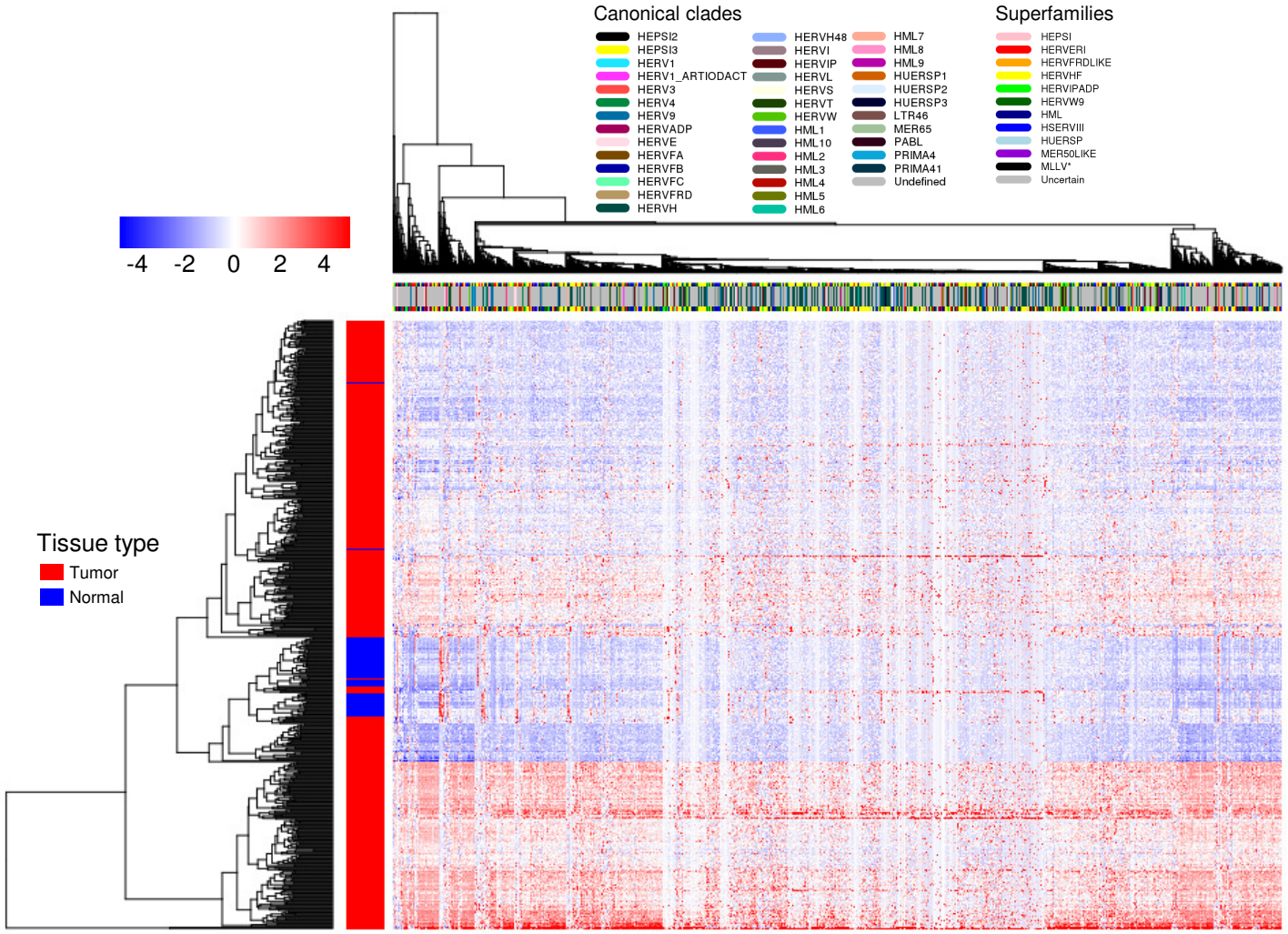


Figure S16: Heatmap of read-normalized hERV expression in TCGA KIRC. Row-side color bar represents tumor (red) and matched-normal (blue) samples. Column-side color bars represent hERV superfamily and canonical clade classifications. Colors are defined by z-score of counts, normalized by each hERV across all samples.

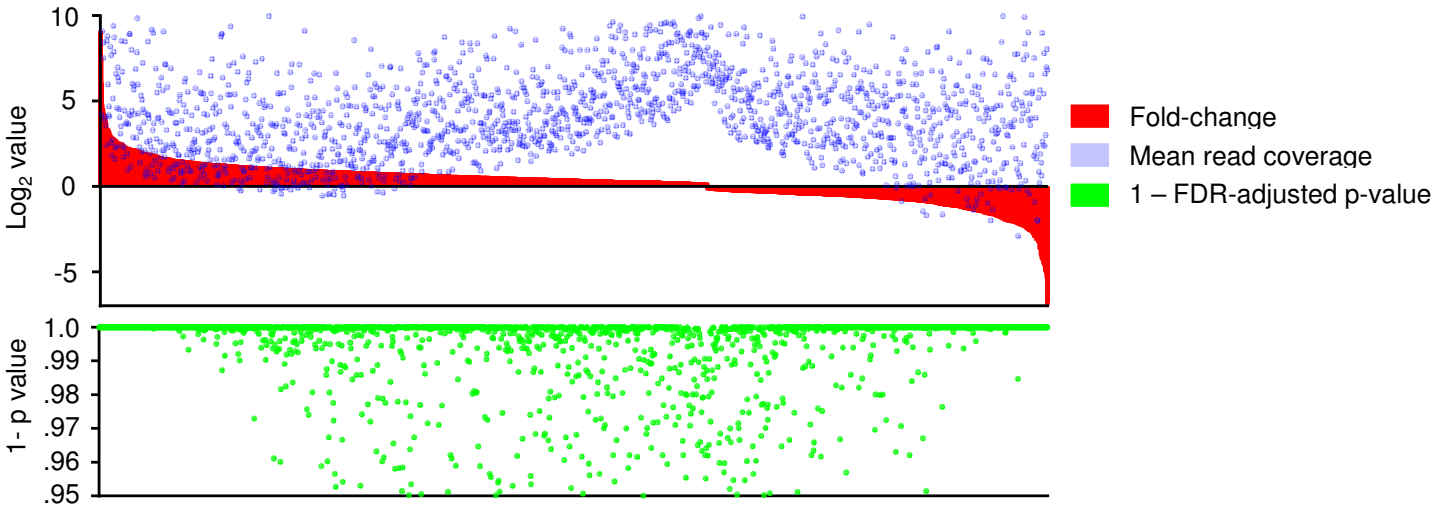


Figure S17: Fold-change expression of differentially expressed hERVs in TCGA KIRC tumor versus matched normal tissue. Data include DESeq2 derived log2(fold change) expression among hERVs with FDR-adjusted p-value ≤ 0.05 (red), mean read coverage (blue dots), and 1 - FDR-adjusted p-value (green dots).

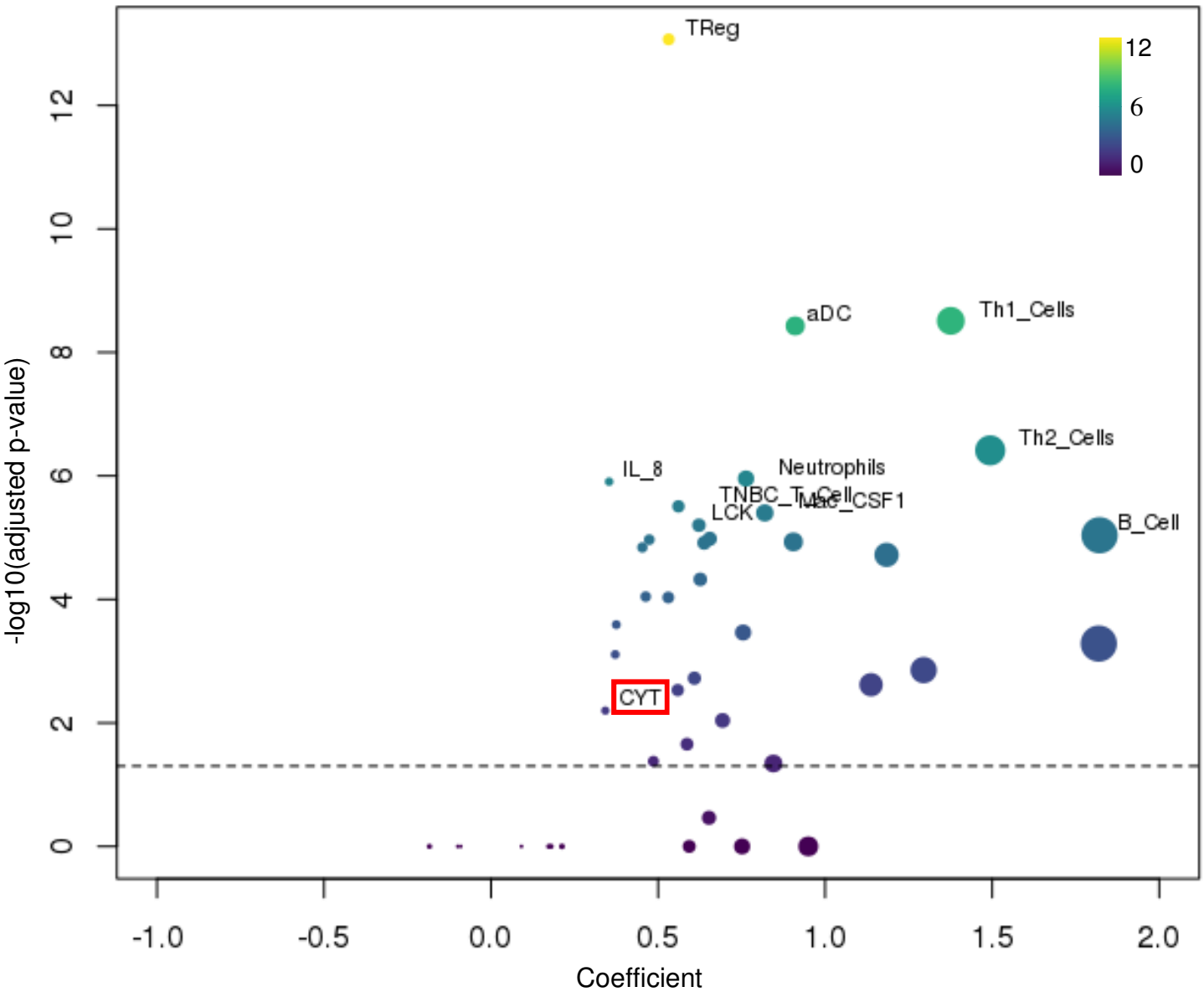
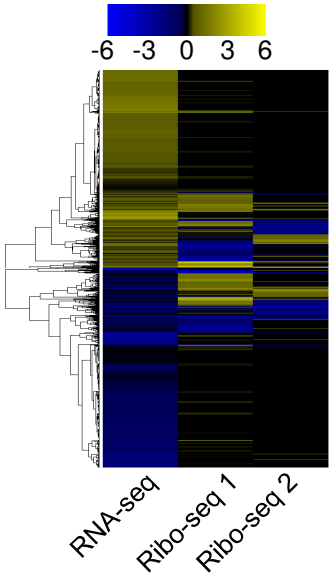


Figure S18: Association between read-normalized hERV expression and immune gene signatures within TCGA KIRC dataset. Graph displays –log10 FDR adjusted p-value (GLM) along the y axis and coefficient along the x-axis. Dashed line represents FDR-corrected p-value = 0.05. Size of each point represents the magnitude of coefficient, and color of each point represents degree of significance with lighter points representing lower p-value.



hERV	RNA-seq log2 fold change	Ribo-seq 1 log2 fold change	Ribo-seq 2 log2 fold change	RNA-seq + average ribo-seq fold change
4700	9.00	0	1.37	10.38
2637	1.33	6.47	1.89	5.51
5875	3.85	3.81	-0.67	5.42
1745	2.25	2.40	3.87	5.38
6169	1.12	5.12	2.79	5.07
3038	3.12	3.18	0.71	5.07
4770	0.73	4.26	0	4.99
2543	0.97	3.65	0	4.61
506	2.09	0	2.51	4.60
5440	3.94	0.61	0	4.54

Figure S19: Differentially expressed hERVs by RNA-seq and Ribo-seq analysis of tumor versus matched normal tissue. DESeq derived heatmap (left) of log2 fold-change for differentially expressed hERVs in TCGA KIRC relative to matched normal tissue by RNA-seq and two independent Ribo-seq analyses of ccRCC tumors, along with quantification of log2 fold-change expression values in the top ten differentially expressed hERVs (right). Top 10 differentially expressed hERVs (right) are ranked by the sum of the RNA-seq fold change expression and the average Ribo-seq fold change expression, filtering for only hERVs with non-zero expression in at least 1 Ribo-seq set. hERV 4700, highlighted in yellow, demonstrated highest differential expression in the tumor as observed in all three datasets.

Figure S20: Read coverage from Riboseq data for CT-RCC hERV-E within the GWIPS database. Tracks represent reads from RiboSeq (red) and RNA-seq (green) datasets. Grey bars along the right-hand side represent aggregate RNA-seq and Ribo-seq data from the entire database ("All"), Ribo-seq data from lymphoblastoid cell line tumor samples ("Tumor-like"), or all other Ribo-seq sets not encompassed by "Tumor-like" ("Other"). All Ribo-seq tracks are linearly scaled between 0 and 186 read coverage.

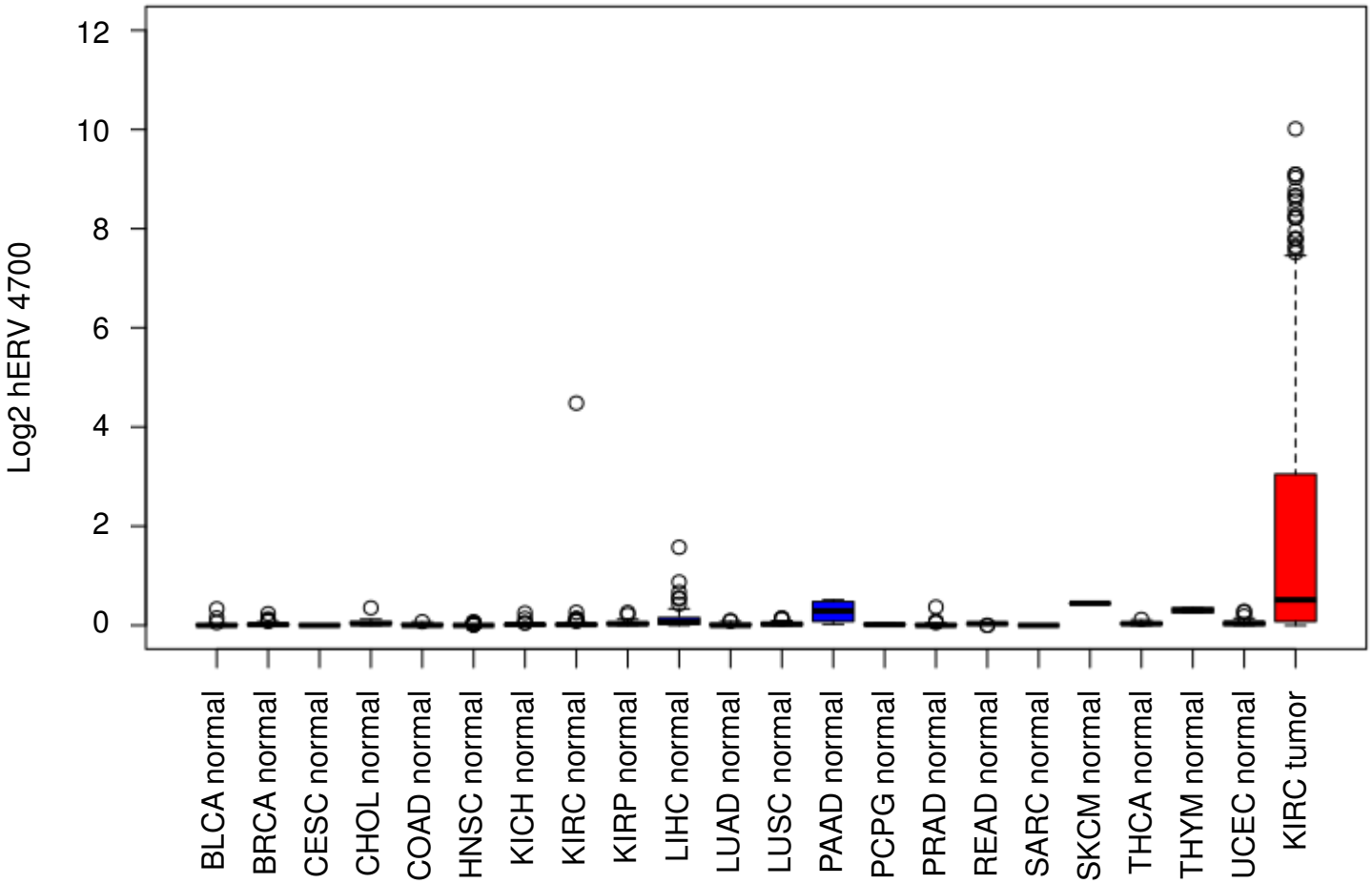


Figure S21: Boxplots of log2 hERV 4700 expression among TCGA pan-cancer matched normal datasets (blue) and TCGA KIRC tumor dataset. Data represent median (middle line), with box encompassing the 25th to 75th percentile, whiskers encompassing 1.5x the interquartile range from the box, and outliers shown by dots.

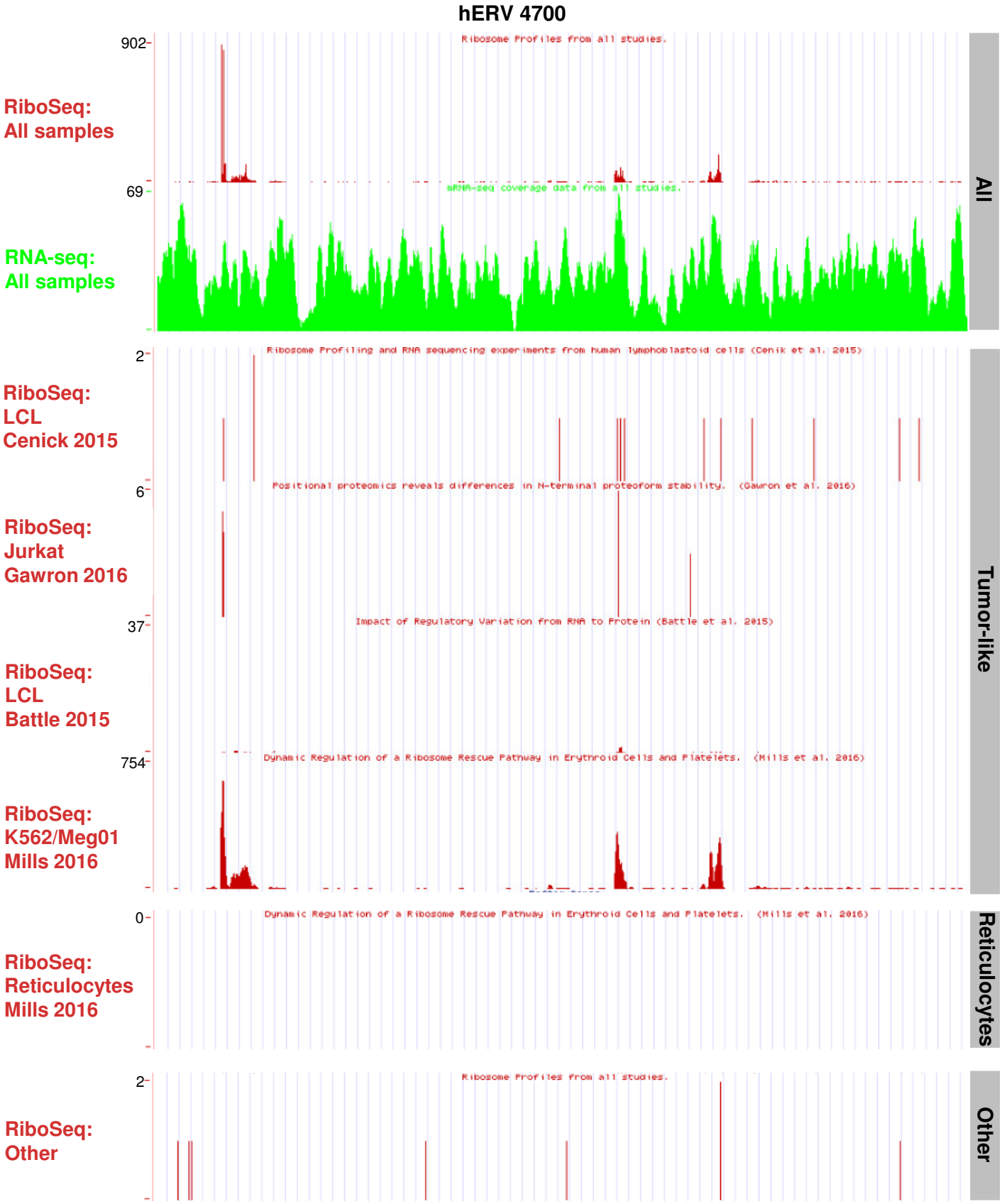


Figure S22: Read coverage from Riboseq data for hERV 4700 within the GWIPS. Tracks represent reads from RiboSeq (red) and RNA-seq (green) datasets. Grey bars along the right-hand side represent aggregate RNA-seq and Ribo-seq data from the entire database ("All"), Ribo-seq data from tumor cell line and lymphoblastic cell line sets ("Tumor-like"), normal primary reticulocytes ("Reticulocytes"), or all other Ribo-seq sets not encompassed by "Tumor-like" and "Reticulocytes" ("Other"). All Ribo-seq tracks are linearly scaled, with the y-axis normalized to the maximum read coverage within each track.

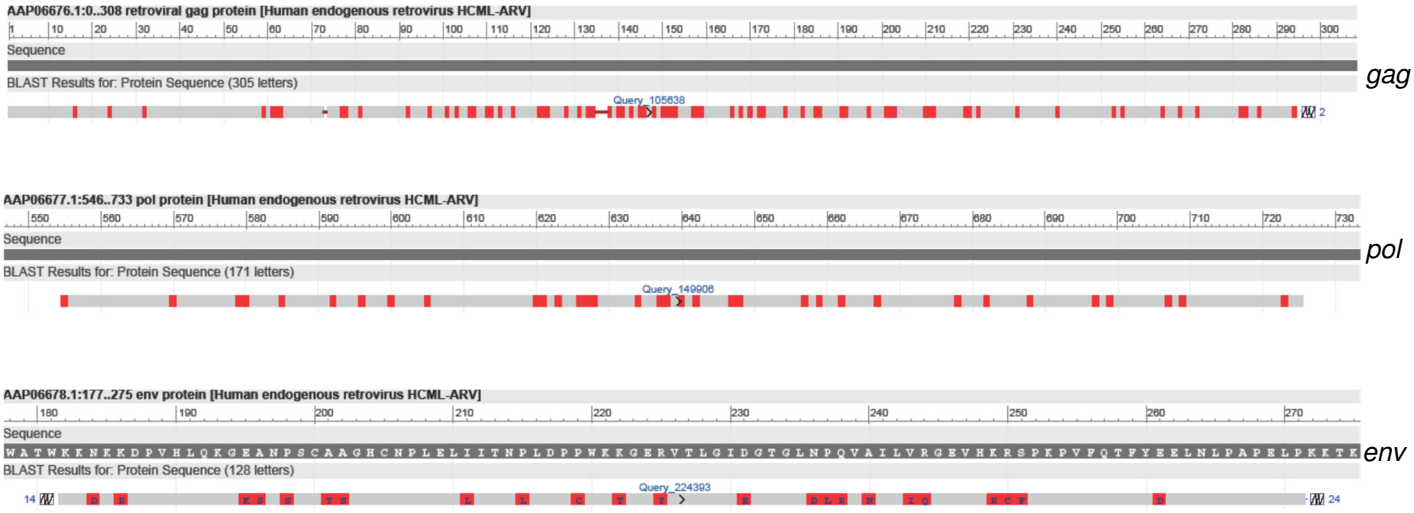


Figure S23: Visual representation of retroviral BLAST results for each of the longest translated *gag*, *pol*, and *env* amino acid sequences for hERV 4700 relative to known retroviral protein sequences. Each track represents the reference sequence (top) with the hERV 4700 translated sequence (bottom), with conserved amino acid positions displayed as grey and non-conversed amino acid positions displayed as red.

Peptide	Protein read-frame	Affinity (nM)	Percent exchange
SLFGKWFPA	env_f2	3.2217	60.65533783
YLSKQLDGV	pol_f2	7.2794	64.07024961
YIDTWLQLV	gag_f2	7.4186	52.5245955
TLCEIDWPA	gag_f2	12.0694	61.63102691
MSLDWELYV	pol_f3	19.2206	57.80957801
AIIDLLQTI	gag_f2	22.9042	67.40385397
YLLATEGGV	env_f3	28.7881	60.49272299
HMVERHAFV	gag_f2	41.0543	62.85063826
SLLCENLCI	pol_f3	53.3321	53.09374746
FLTLQVHGA	pol_f3	53.875	64.15155704
LLKEQDIPL	pol_f1	87.9206	73.09537361
QLLMYLFNM	gag_f2	92.9628	58.62265225
ALGGFKTLV	env_f2	94.4266	65.12724612
FVYQPFNAA	gag_f2	106.9197	18.53809253
YTDSQYAFL	pol_f3	116.1574	60.98056753
MVLRLDVPL	pol_f1	125.3523	54.55728108
RLQAILEII	env_f3	136.2401	50.49190991
GVLPLIPTA	gag_f3	167.6426	58.94788194
LLPVSESPV	pol_f3	230.4273	59.11049679
MVGPWPRPV	pol_f2	238.161	64.47678673
NSWQEMVPV	gag_f2	245.9918	62.11887145
NLLDPCWKT	env_f1	260.5415	42.68639727
VLPLIPTAL	gag_f3	290.1308	16.09886983
YAVVTLDAV	pol_f3	290.989	58.13480771
LVLGPPWWL	gag_f2	296.4327	21.13993008
NLTNCCLQI	env_f3	334.0793	39.02756322
VLMAKGQTA	gag_f2	352.4522	41.87332303
TLVIGIIIV	env_f2	413.3257	28.70152045
RLALDYLLA	env_f3	420.0978	17.56240345
RLTRYQSLL	pol_f3	465.0925	57.80957801

Table S5: NetMHCPan4.0 results and HLA-A*02:01 monomer UV exchange efficiency for peptide sequences identified by translation of proviral sequences with both RNA-seq and Ribo-seq coverage in hERV 4700. DNA sequences with evidence of coverage by both RNA-seq and Ribo-seq within the hERV 4700 proviral sequence were translated in three reading frames and peptides ≥8 amino acid residues in length were submitted to NetMHCPan4.0 for MHC-binding prediction to HLA-A*02. Table displays the peptide sequence, reading frame, predicted peptide-MHC binding affinity, and HLA-A*02:01 monomer exchange efficiency.

Sample ID	Histology	Sex	Age at time of IO initiation	IO Agent	Best Response	Duration on Therapy	Race
NR_1	clear cell	M	46	Nivolumab	Progressive Disease	12 weeks	White
NR_2	clear cell	M	60	Nivolumab	Progressive Disease	8 weeks	White
NR_3	clear cell	M	49	Nivolumab	Progressive Disease	12 weeks	White
NR_4	clear cell	M	79	Nivolumab	Progressive Disease	18 weeks	White
NR_5	clear cell	M	54	Nivolumab	Progressive Disease	10 weeks	Hispanic
NR_6	clear cell	F	63	Nivolumab	Progressive Disease	8 weeks	White
R_1	clear cell	M	72	Nivolumab	Partial Response	11 months	White
R_2	clear cell	M	67	Nivolumab	Partial Response	7 months	White
R_3	clear cell	M	54	Atezolizumab	Partial Response	28 months	White
R_4	clear cell	M	48	Nivolumab	Partial Response	>24 months	White
R_5	clear cell	M	63	Nivolumab	Partial Response	8 months	White
R_6	clear cell	F	63	Nivolumab	Partial Response	9 months	White
R_7	clear cell	F	73	Nivolumab	Partial Response	13 months	White

Table S6: **Patient demographic information from anti-PD-1 treated ccRCC samples represented in RT-qPCR data in Figure 5E.**

	Forward primer (sense)	Probe (sense)	Reverse primer (anti-sense)
gag 1	GACGCTCCCAGCAGAATAAA	TTGTCTGTGGCTTGTCCTGCTACA	CCGGTCAGGAAACCAAGAAA
gag 2	GTCCTGCTACATTTCTTGGTTTC	TGATTAAGGGACAGTGGAGGCAGC	GCACTCTCAGGATCCACATT
pol 1	GTGTGGGATATGCAGTGGTAA	TGTCATTGAAGCCAAATCGTTGCCC	GGCCCGAATTAAAGCAATGAG
pol 2	CCATCCTTGGATGTCACTAGAC	TACGTGGACGGGAGCAACTTTGTC	CCAGGGTTACCACTGCATATC
env 1	CTGCTTAGGTCCATCCAGAATC	ACGGCTCCCTCTGGACTATACTGG	TGATCAGGTGACGGAGTGTA
env 2	CCAGGCCTGTAGGTTAAAGATT	CCCAACCGCTTGTGCTATCCATAGA	GTGGTGAGGAAGGCAAGTATT

Table S7: Primer and probe sets for hERV 4700 *gag*, *pol*, and *env* RT-qPCR assay.

Sample ID	*gag* 1	*gag* 2	*pol* 1	*pol* 2	*env* 1	*env* 2	Response		
NR_1	678.4063448	NA	471.9946807	1.05076425	1196.291466	45.59265581	N		
NR_2	1966.48594	524.4147042	1477.437437	668.5377752	2893.493623	122564.4228	N		
NR_3	0.124258101	8.443259534	44.70848193	27.53499411	21.38764759	176.4189823	N		
NR_4	2.871319227	1798.599326	499.7381473	0	0	0	N		From patient R_1
NR_5	385.4097904	0	280.4108268	181.1442066	104.0998151	488.0130608	N		From patient R_4
NR_6	565.3428135	338.0681755	1902.726334	456.6574505	188.2126195	2069.477061	N		
R_1.1	5258.936393	1230.618198	161.1495075	1255.96081	12990.86869	2771.044374	Y		
R_1.2	246545.7566	32137.56664	125544.7762	65993.11381	319483.7794	118531.3066	Y		
R_2	55891.70575	9974.190645	28255.53299	8984.664465	75429.48579	38010.33345	Y		
R_3	951048.0824	360566.493	721664.382	152893.0483	626782.8396	265134.4676	Y		
R_4.1	300.7759069	75.08179676	185.2933803	30.75194966	40.3559261	474.8349328	Y		
R_4.2	17697.77791	8741.174546	35544.39322	12770.52268	11043.52628	152347.8539	Y		
R_5	13345.98475	223.3670108	2232.037754	2369.17165	8604.665228	559331.3482	Y		
R_6	5177.153134	407.9995433	1818.74592	1169.360258	3523.835961	187125.7868	Y		
R_7	45716.73169	11738.80477	18013.19592	11446.60486	24849.38302	1252395.658	Y		

Table S8: hERV 4700 *gag*, *pol*, and *env* RT-qPCR non-transformed expression values.

qPCR region	Mann-Whitney p-value
gag 1	0.0012
gag 2	0.0338
pol 1	0.0023
pol 2	0.0023
env 1	0.0047
env 2	0.0082

Table S9: Mann-Whitney u-test p values for comparison of *gag*, *pol,* and *env* specific RT-qPCR primer/probe set signals in non-responder (n = 6) and responder (n = 7) tumor samples from anti-PD-1 treated ccRCC patients. Data represent statistical testing of RT-qPCR experiments shown in figure 5E.

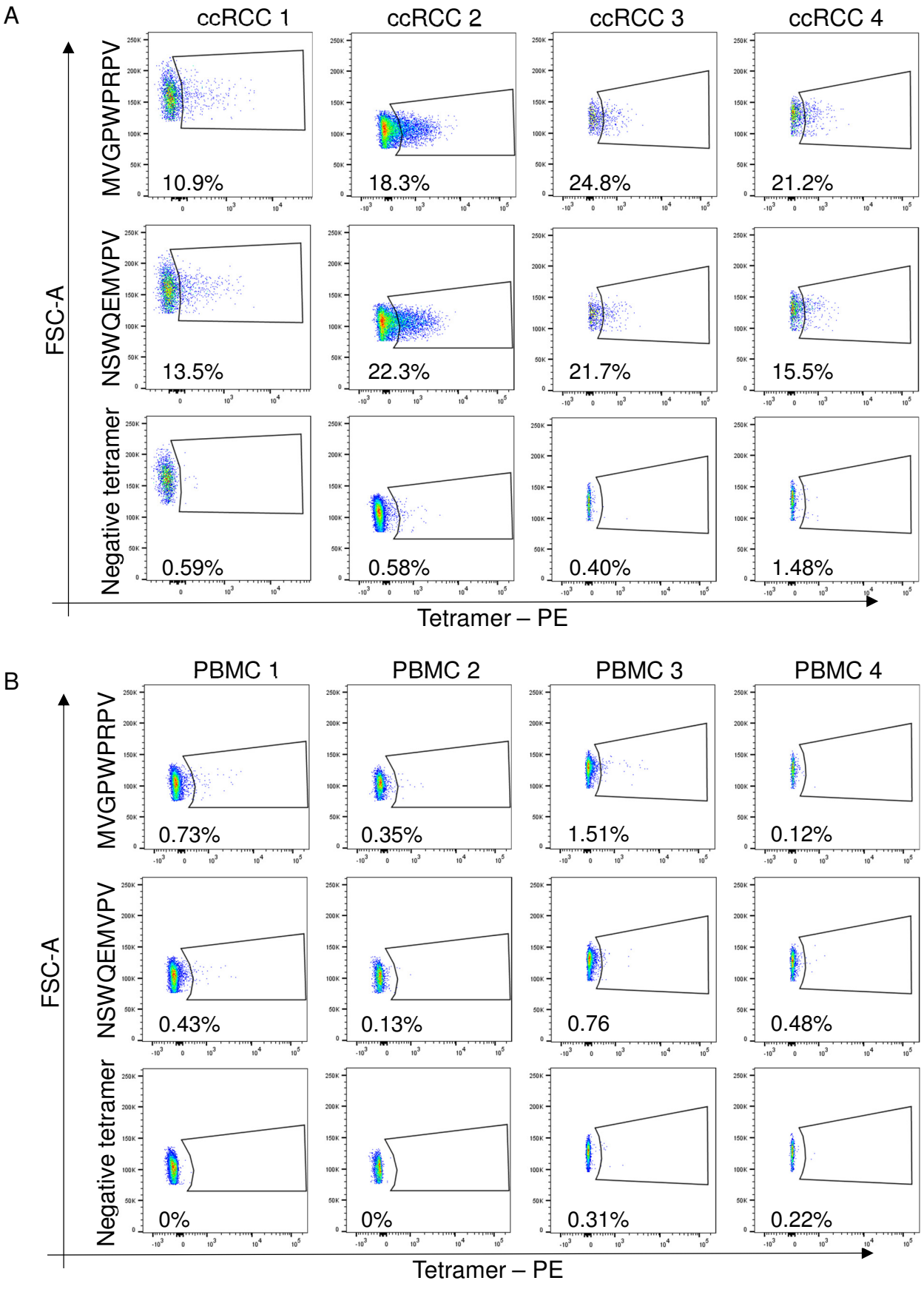


Figure S24: hERV 4700 epitope tetramer staining. Flow cytometric analysis in (A) ccRCC tumors (n = 4) and (B) healthy donor PBMCs (n = 4), staining for HLA-A*02:01 tetramers containing hERV 4700 epitopes within *gag* (MVGPWPRPV, Figure 6B tetramer 2) and *pol* (NSWQEMPV, Figure 6B tetramer 3) proteins or with a negative control tetramer. Samples are gated according to representative gating shown in Figure 6A. Data represent results from three independent experiments.

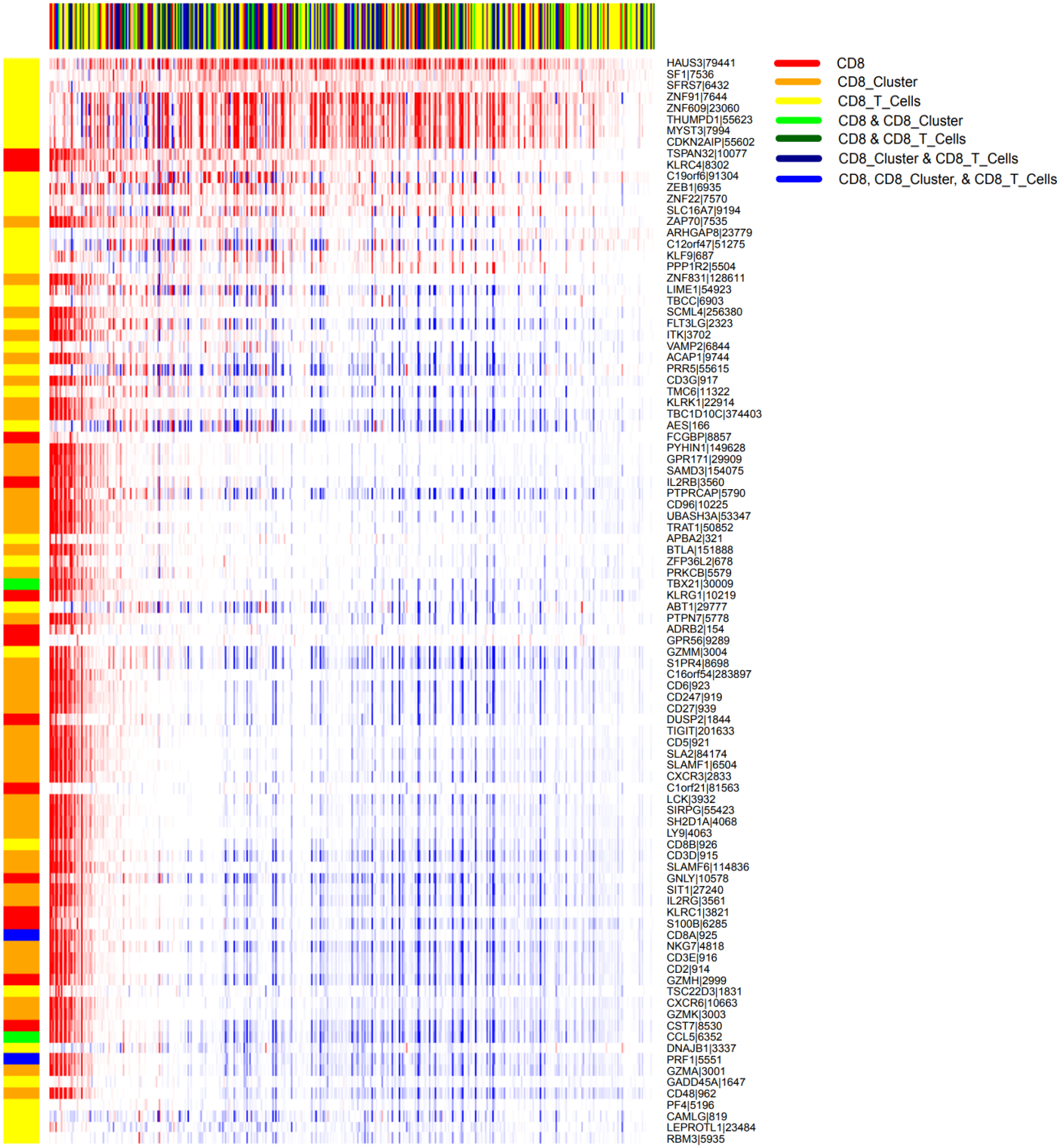


Figure S25: Heatmap of association between hERV expression and genes from three CD8 T cell IGS among TCGA pan-cancer dataset. FDR corrected p-values (GLM) represented by intensity of color and direction of coefficient represented by color (red: positive, blue: negative). Row-side color bar represents individual CD8 signature (or signatures) within which each gene belongs. Rows and columns are ordered by number of significantly positive associations.

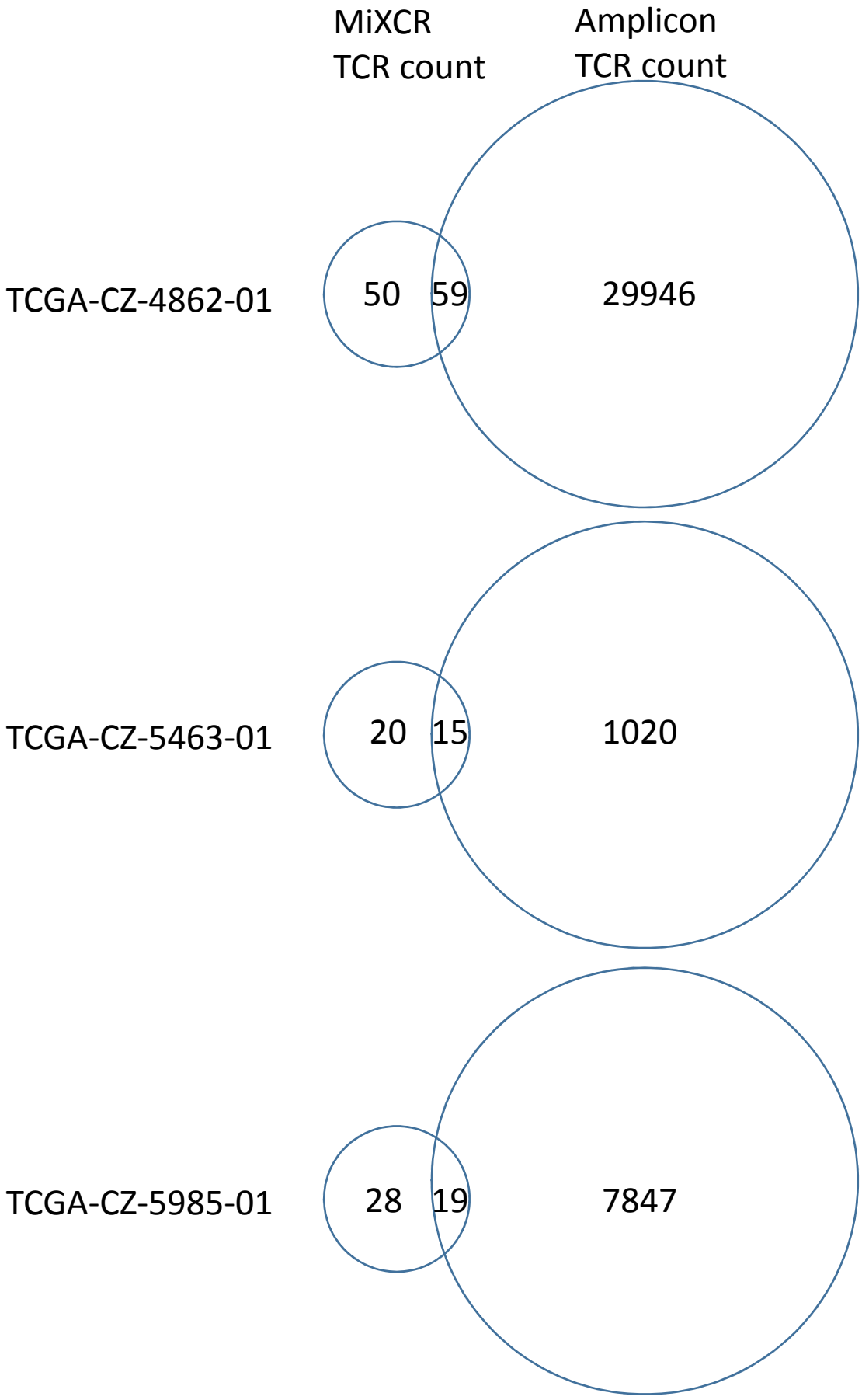


Figure S26: Number of unique T cell receptor clones identified in three TCGA KIRC samples by MiXCR-based RNA-seq TCR inference or Adaptive TCR amplicon profiling.

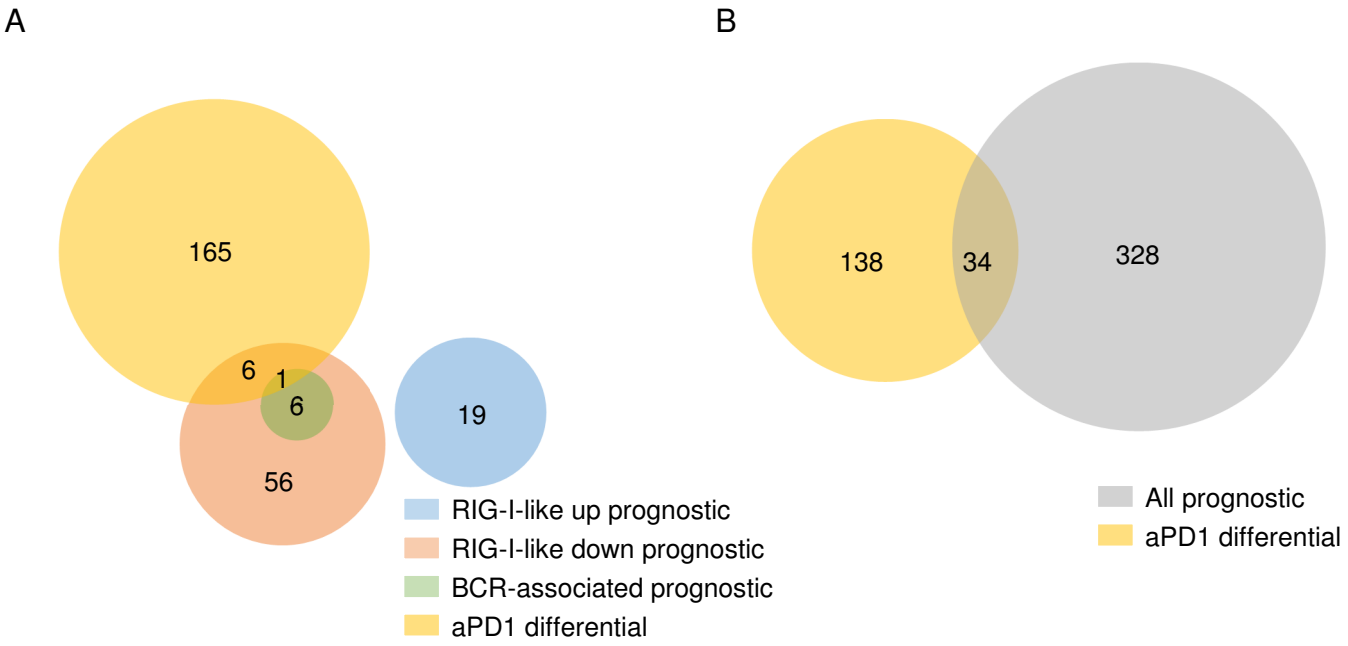


Figure S27: Overlaps between hERV signatures and groups. (A) Venn diagram of hERVs within prognostic (RIG-I-like up/down and BCR-associated) signatures and differentially expressed hERVs between aPD1 responsive versus non-responsive ccRCC tumors (Mann-Whitney $p < 0.05$). (B) Venn diagram of total prognostic and differentially expressed hERVs in ccRCC.

Tables S10-S14
Too large to display

Table S10: Summary of CoxPH analysis with read-normalized hERV expression as a predictor for overall survival. Data are displayed as significance index ($-\log_{10}$(p-value) – $\log_{10}(0.05)$ * Direction of coefficient).

Table S11: Summary of CoxPH analysis with UQN hERV expression as a predictor for overall survival. Data are displayed as significance index ($-\log_{10}$(p-value) – $\log_{10}(0.05)$ * Direction of coefficient).

Table S12: Read-normalized hERV expression matrix for TCGA pan-cancer dataset.

Table S13: UQN hERV expression matrix for TCGA pan-cancer dataset.

Table S14: Raw hERV expression matrix for anti-PD-1 treated ccRCC tumor samples.