

Supplemental information

**Single-cell RNA-seq-based proteogenomics
identifies glioblastoma-specific transposable
elements encoding HLA-I-presented peptides**

Pierre-Emmanuel Bonté, Yago A. Arribas, Antonela Merlotti, Montserrat Carrascal, Jiasi Vicky Zhang, Elina Zueva, Zev A. Binder, Cécile Alanio, Christel Goudot, and Sebastian Amigorena

Supplementary 1

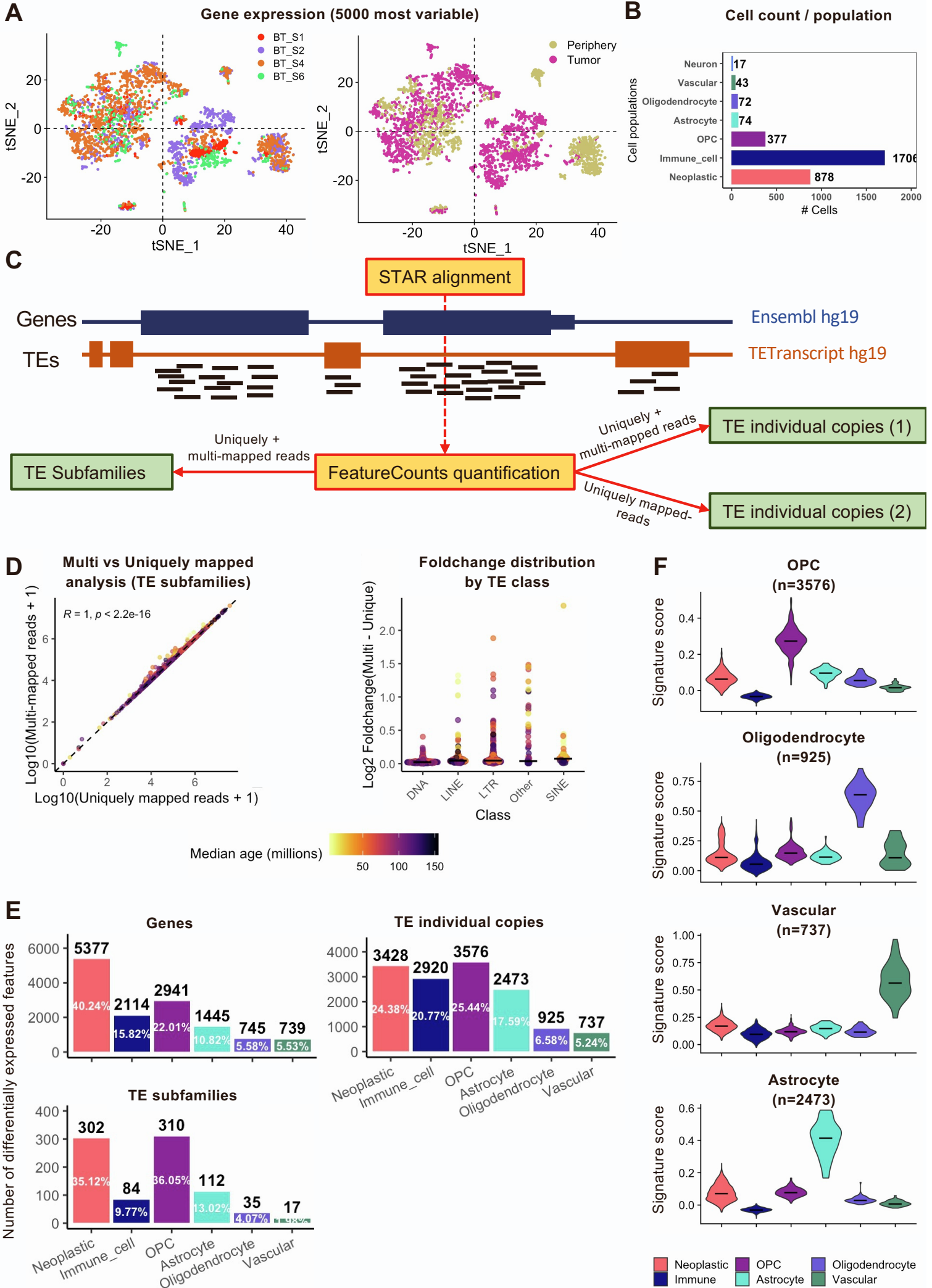
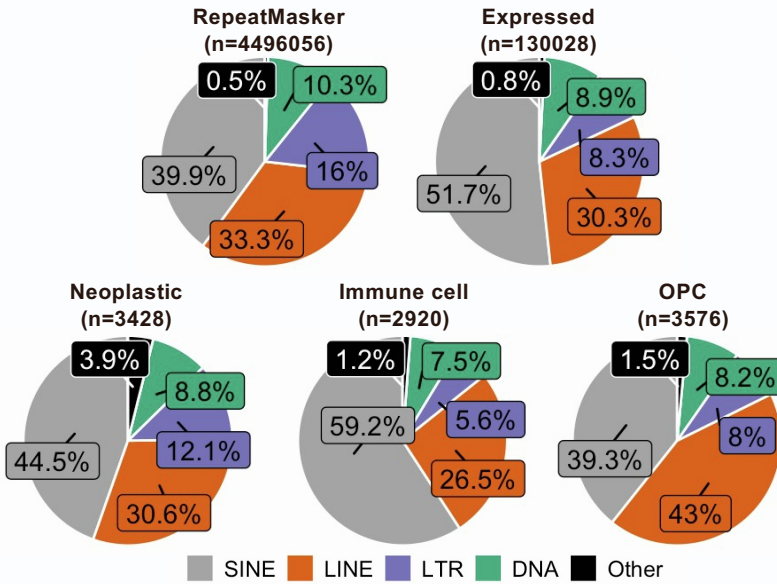


Figure S1. General description of single cell dataset, workflow, genes and TE signatures, Related to Figure 1. (A) t-Distributed Stochastic Neighbor Embedding (tSNE) visualizing all single cells after filtering ($n = 3,167$), colored by patient ID (left) or location (tumor core and surrounding tissue, right). (B) Barplot showing the number of cells in each cell population. (C) Workflow showing the strategy of alignment and TE quantification using uniquely or multiple mapped reads. (D) On the left, plots displaying the correlation of expression in each subfamily between the quantification with uniquely mapped reads (x-axis) and multiple mapped reads (y-axis). On the right, the log₂foldchange between multiple mapped reads vs uniquely mapped read quantification. Each dot represents a TE subfamily and a scale color is used to show the median of age of TEs within each subfamily. (E) Barplots showing the number of differentially expressed genes (top left), TE subfamilies (bottom left) and TE individual copies (right) in each cell population. (F) Violin plots representing the signature score for OPC, Astrocyte, Oligodentocyte, and Vascular cell populations based on their differentially expressed TEs. The number of differentially expressed TEs included in each signature is indicated.

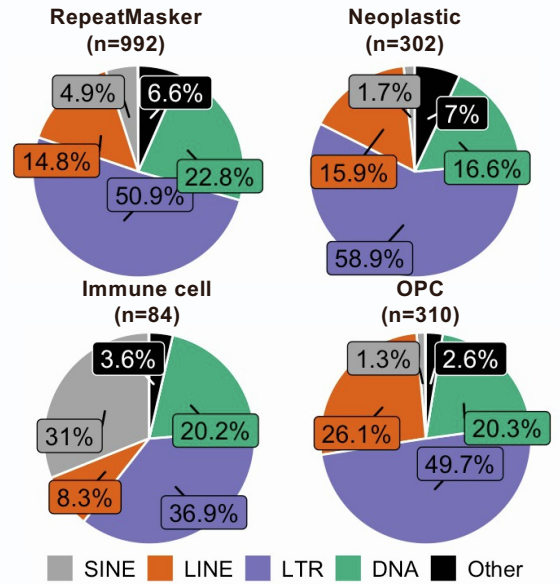
Supplementary 2

A

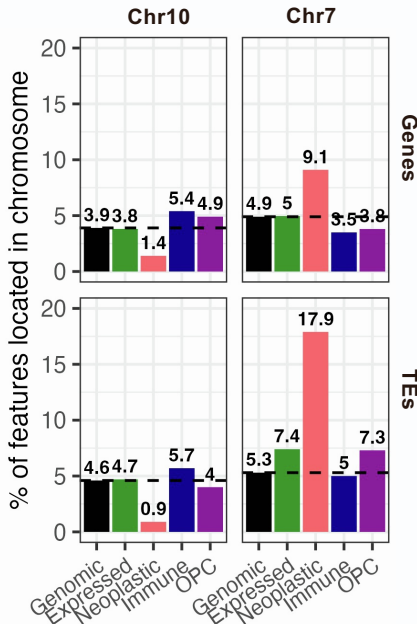
Class distribution in individual TE copy signatures



B Class distribution in TE subfamily signatures

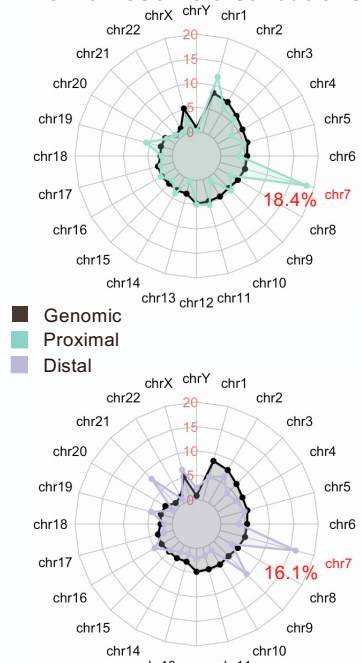


C Proportions of TEs and genes per chromosome



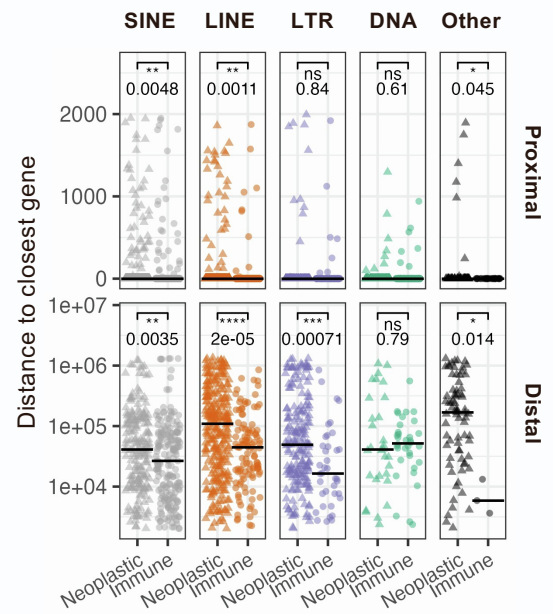
E

Neoplastic TE chromosome distributions

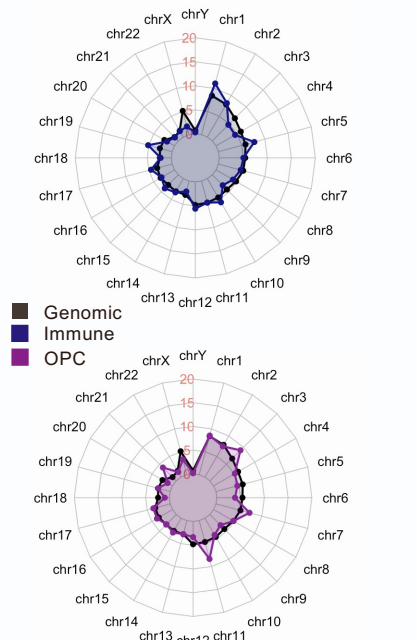


G

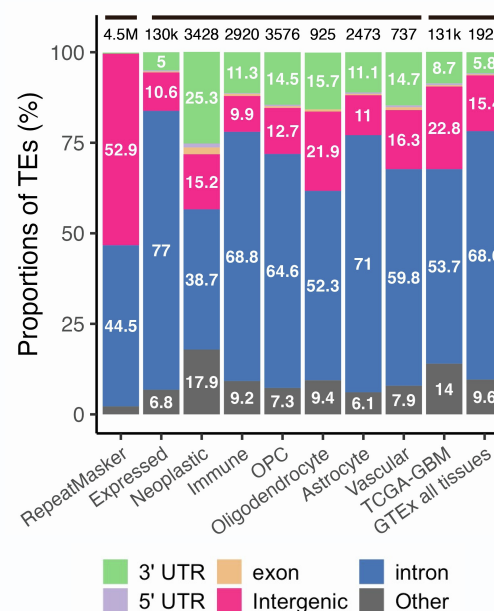
Distance to closest gene distribution in individual TE copy signatures



D TE chromosome distributions



F TE genomic location distributions



H

Gene-Individual TE copy correlation

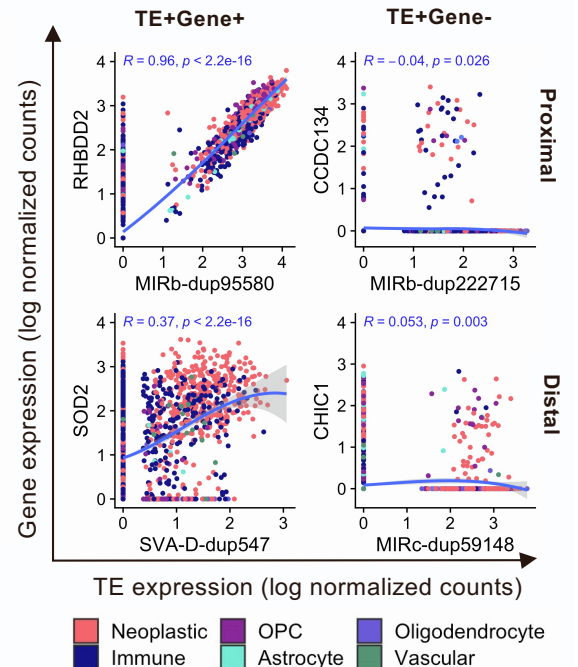
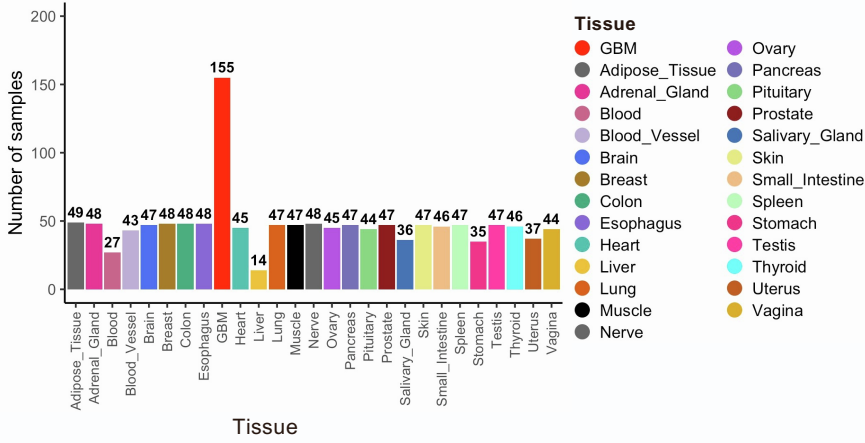


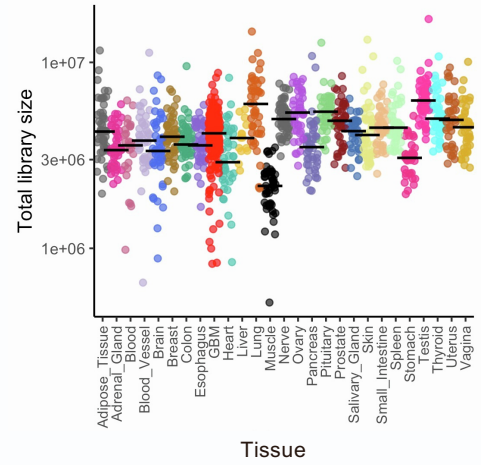
Figure S2. TE class, chromosome and genomic location distributions observed among single cell signatures, Related to Figure 1. (A-B) Pie charts showing TE class distributions within individual TE copy (A) or TE subfamily signatures (B) as compared to class distribution observed in the genome (RepeatMasker) or in all filtered expressed TEs (Expressed). (C) Barplots showing the rate of genes (first line) or TEs (second line) located in chromosome 10 (left) or 7 (right) on different subsets of features: All annotated features in the genome (Genomic), all expressed features in the data set after filtering (Expressed), all differentially expressed features from neoplastic, immune and OPC cell populations. (D) Radar plots displaying the rate of TEs along all chromosomes for immune cells (top) and OPC cells (bottom). Genomic distribution from RepeatMasker is plotted in black. (E) Radar plots displaying the rate of proximal (top) or distal (bottom) neoplastic TEs along all chromosomes. Genomic distribution from RepeatMasker is plotted in black. (F) Barplot showing the distribution of different types of RefSeq genomic locations for individual TE copies within RepeatMasker, expressed TEs in all cell populations, TE signatures for each cell population and TEs expressed in bulk RNA-seq TCGA-GBM and GTEx data sets. (G) Plot showing the distance to closest protein-coding gene per class of TEs for proximal (first line) and distal (second line) TEs comparing neoplastic and immune TE-signatures. (H) Scatter plots illustrating the correlation between TE expression and their nearest genes. The TE⁺gene⁺ category (left) represents a positive correlation when the TE and gene are both differentially expressed. The TE⁺gene⁻ category (right) represents a negative correlation when the TE is differentially expressed and not the gene. The categories are also separated according to proximal (top) and distal status (bottom).

Supplementary 3

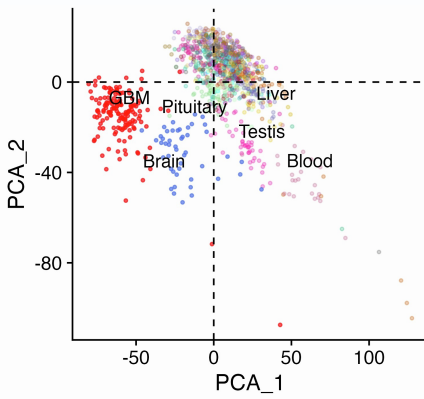
A Number of processed samples per tissue



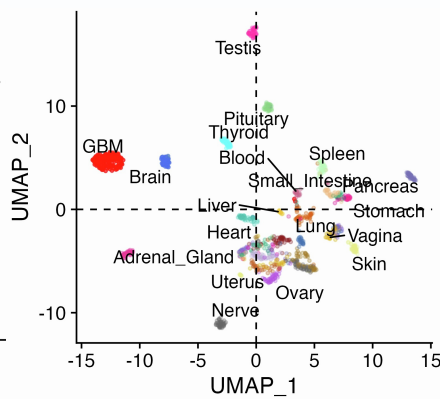
B TE library size distribution per tissue



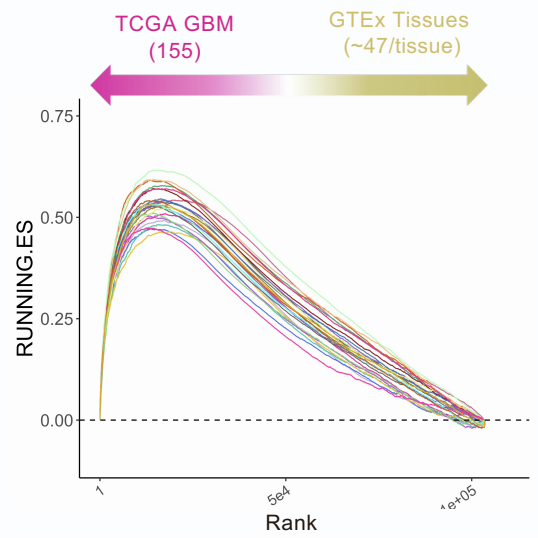
C PCA on TE Neoplastic signature



D UMAP on TE Neoplastic signature



E



F

Tissues	FDR	NES
Adipose_Tissue	0.012	1.76
Adrenal_Gland	0.018	1.71
Blood	0.0039	1.87
Blood_Vessel	0.028	1.71
Brain	0.027	1.67
Breast	0.0068	1.77
Colon	0.0033	1.93
Esophagus	0.0082	1.82
Heart	0.025	1.69
Liver	0.029	1.66
Lung	0	1.93
Nerve	0.0097	1.7
Ovary	0.036	1.66
Pancreas	0.02	1.75
Pituitary	0.034	1.64
Prostate	0.0025	1.87
Salivary_Gland	0.0051	1.75
Skin	0.0036	1.73
Small_Intestine	0.0013	1.86
Spleen	0	1.85
Stomach	0.0047	1.8
Testis	0.039	1.6
Thyroid	0.023	1.68
Uterus	0.0075	1.77
Vagina	0.0074	1.76

G

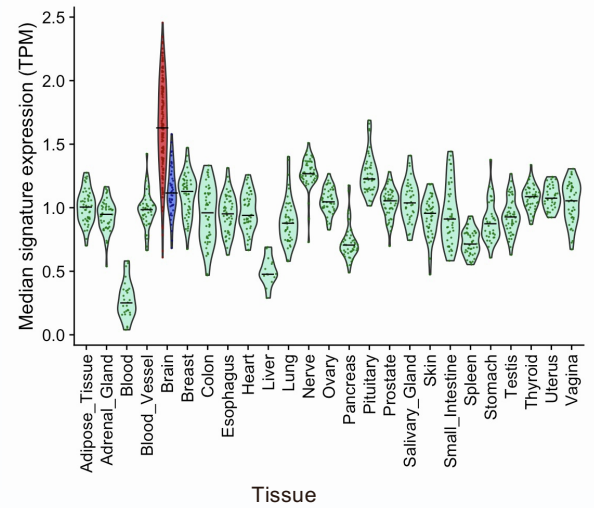


Figure S3. Validation of neoplastic TE-signature using GTEx and TCGA cohorts, Related to Figure 2. (A) Barplot showing the number of samples integrated in each condition (B) Plot representing the TE library size per data set types and tissues. (C-D) PCA and UMAP projection of TCGA. GBM tumor samples and healthy tissue samples from GTEx based on single cell neoplastic TE signature. Each point corresponds to a sample. Samples are color-coded by their tissue of origin. (E) Gene Set Enrichment Analysis (GSEA) was performed to assess the specific enrichment of the neoplastic TE-signature in TCGA GBM tumor samples compared to samples from 25 GTEx normal tissues. (F) Table showing the Normalized Enrichment Score (NES) and FDR for each tissue compared to GBM tumor. (G) Violin plots showing the median expression of single cell neoplastic signature in TCGA-GBM tumor samples and GTEx normal samples.

Supplementary 4

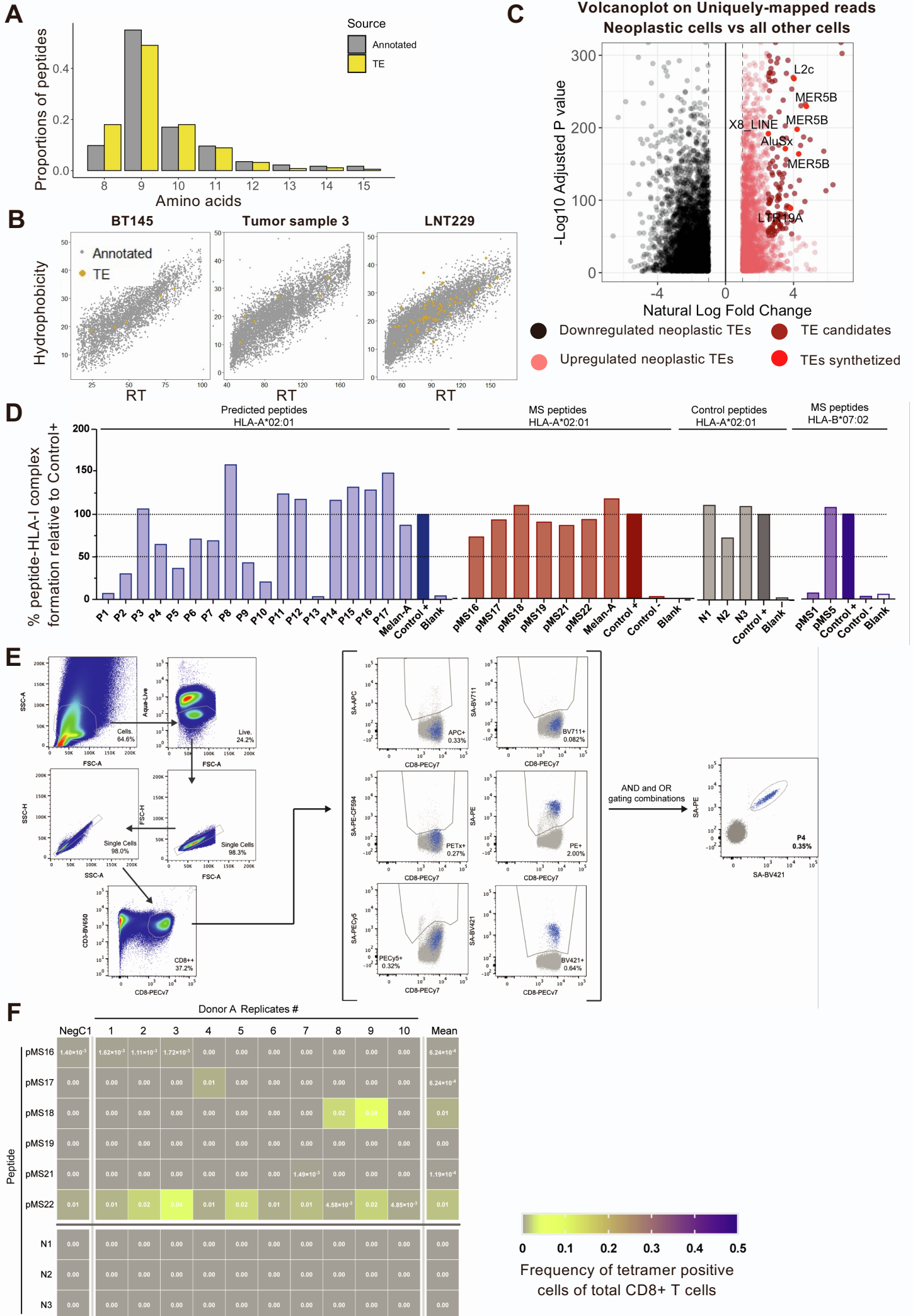


Figure S4. Quality control, strategy and validation of TE-derived peptides, Related to Figure 3. (A) Peptide length distribution (in amino acids) from annotated and TE-derived peptidomes. (B) Scatter plots comparing retention time (in minutes) and hydrophobicity index from annotated and TE-derived identifications. (C) Volcano plot displaying differential TE expression between neoplastic cells vs other cells. Up-regulated (pink) and down-regulated TEs (black) are indicated. TE candidates with low p-value (less than $1e-50$) and high natural log fold change (more than 2) are indicated in dark red. TE synthesized into peptide for immunogenicity study are colored in red. (D) Binding to HLA-A02*01 and HLA-B*07:02 measured as percentage of peptide-HLA-I-complex formation compared to positive control. (E) CD8-tetramer + cells gating strategy. (F) Example of tetramer frequency analysis after *in-vitro* immunogenicity assays. Donor A and HLA-A*02:01 peptides are shown as representative example. Tetramer positive frequencies per replicate (columns) for each evaluated peptide (rows) are indicated. NegC1: negative control replicate with no peptide pulsed on the mo-DCs. Mean: total frequencies of tetramer positive populations for each peptide considering all CD8+ T cells evaluated in all replicates.

Supplementary 5

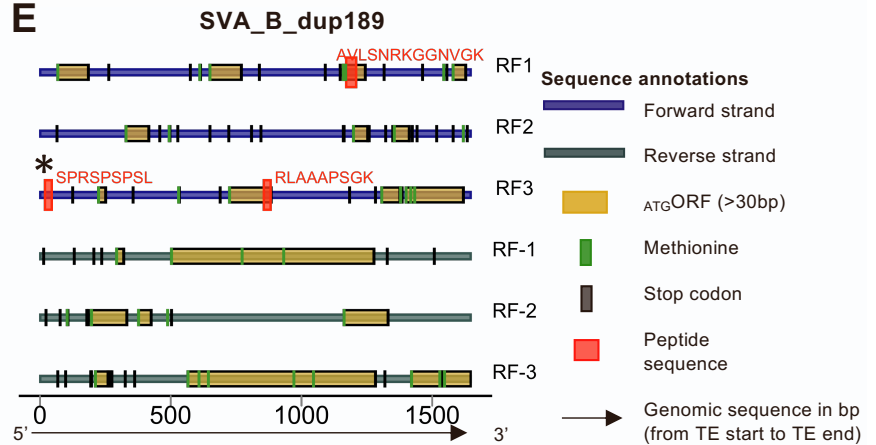
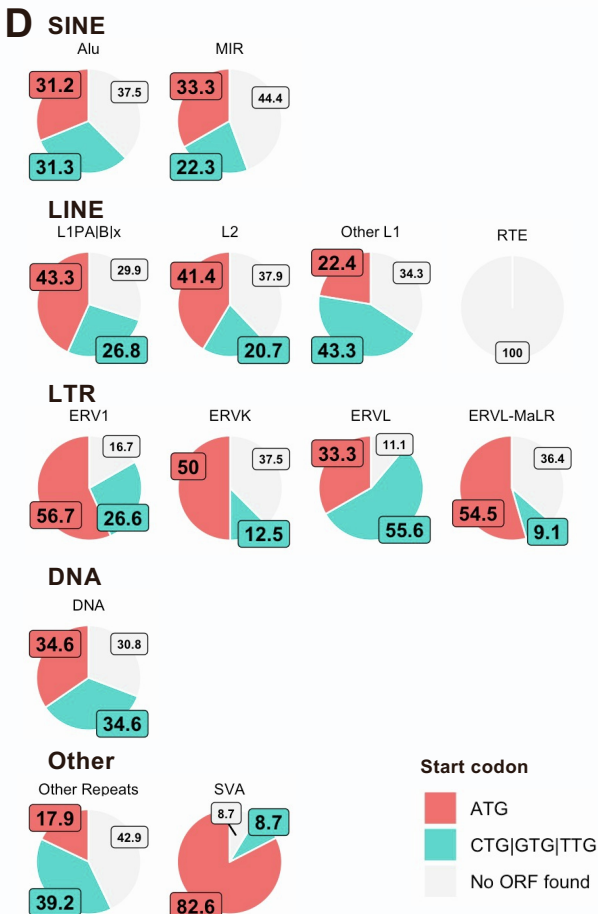
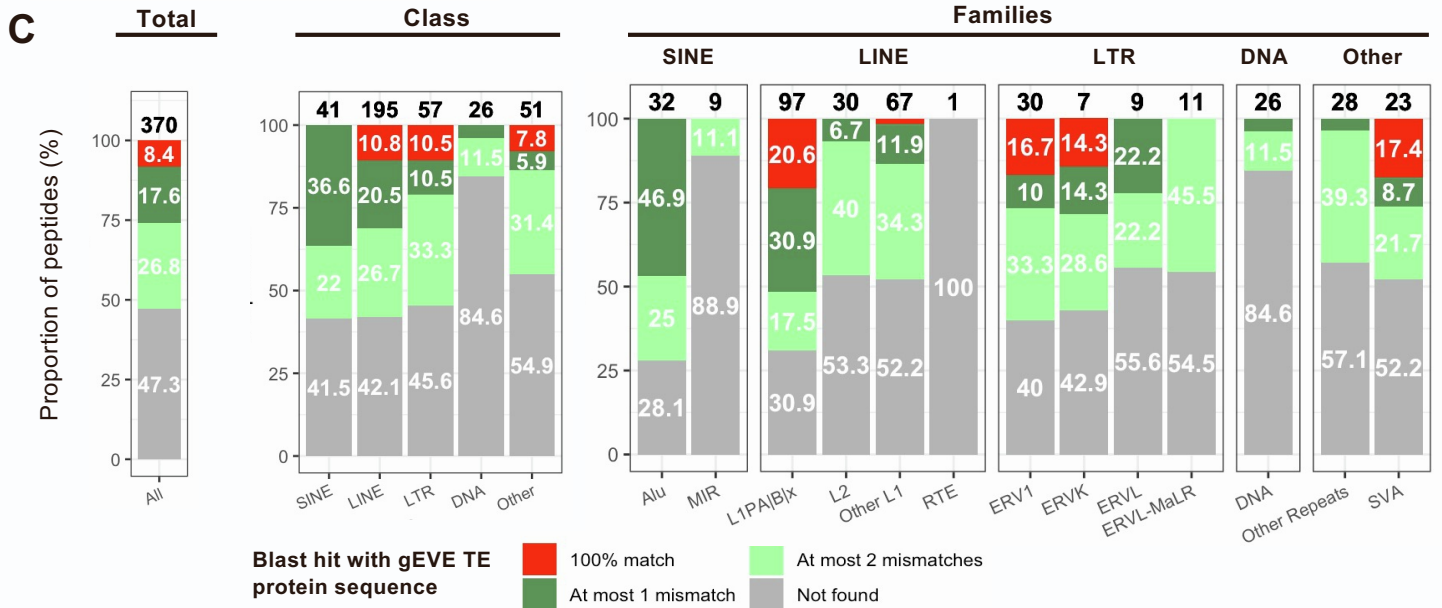
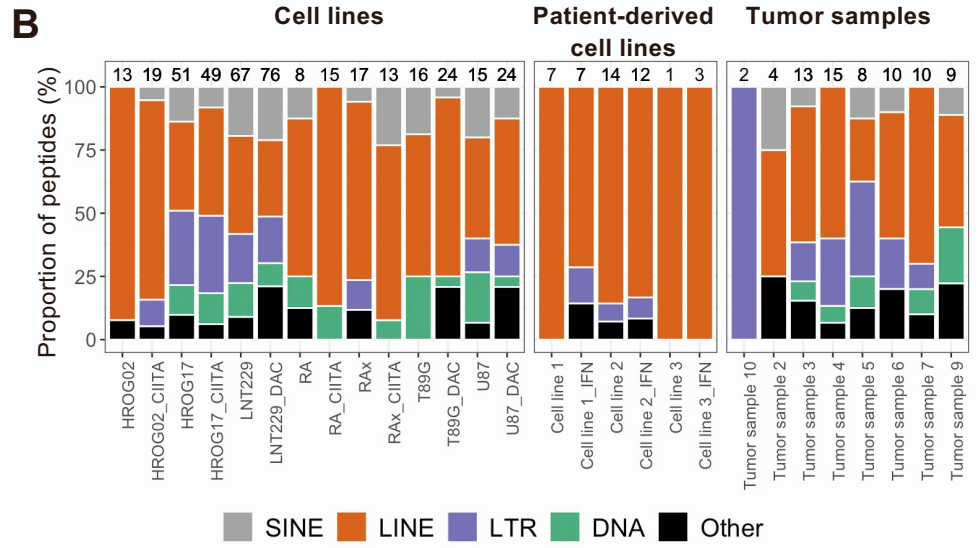
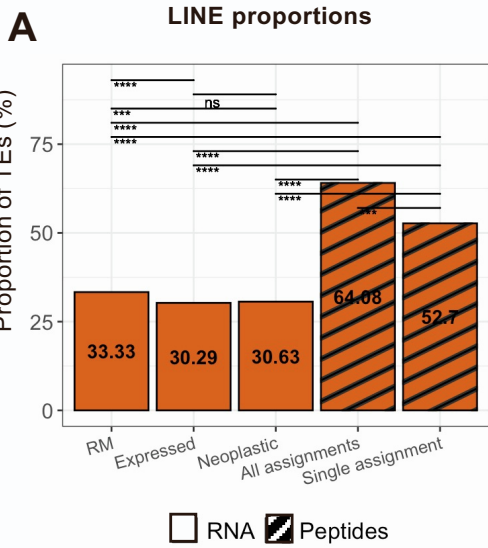
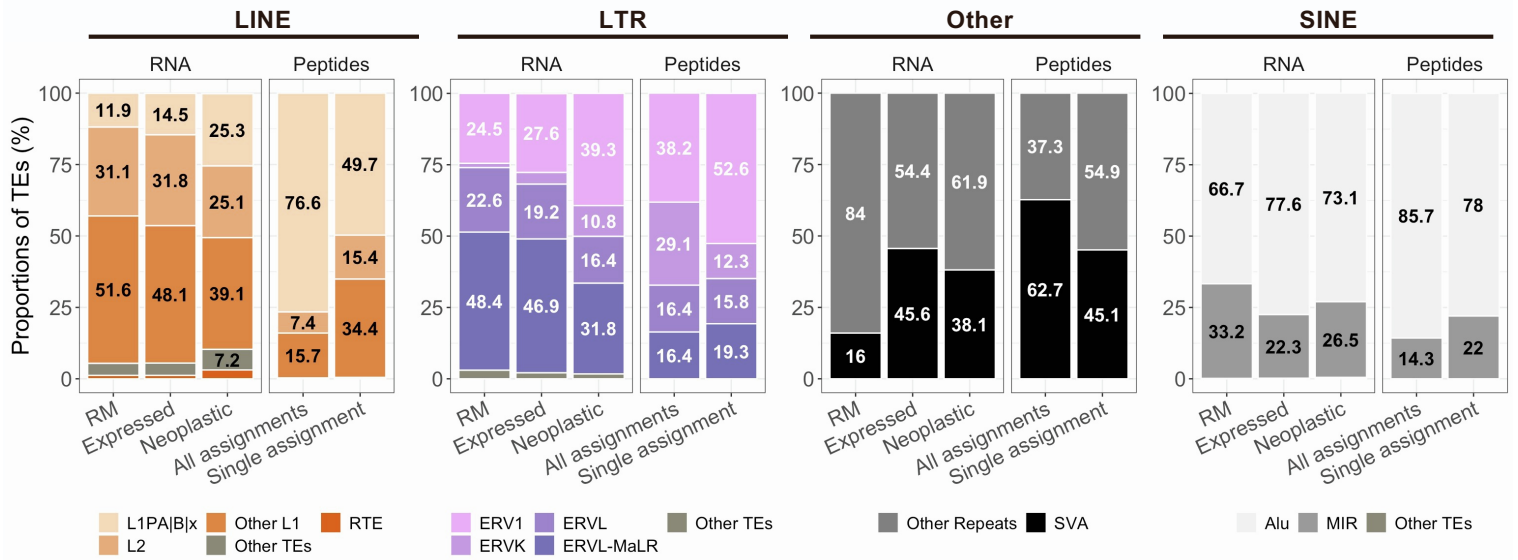


Figure S5. Characterization of TE class and genomic origin of TE-derived peptides, Related to Figure 4. (A) Barplot showing the LINE proportions at RNA level in RepeatMasker (RM), expressed TEs, TEs from neoplastic TE-signature and at peptide level using single or all assignments. (B) Barplots representing among TE-derived peptides the proportions of TE classes per sample used in immunopeptidomics. The samples were classified according to their origin (cell lines, patient-derived cell lines and tumor samples). (C) Barplots showing the percentage of peptides bearing an Endogenous Viral Element ORF documented in the gEVE database using 0 (100% match), 1 or 2 mismatches in all peptides (left panel), peptides grouped by class (middle panel) or peptides classified by TE family (right panel). (D) Pie charts showing the percentage of TE-derived peptides with canonical (red) and non-canonical codon start (blue) per class and TE family (E) Example of one peptide-coding TE: SVA_B_dup189. All reading frames (RFs) (forward strand RF1,2,3; reverse strand RF-1,-2,-3) are represented. Start codons (ATG), stop codons, identified peptides are indicated in green, black and red rectangles, respectively. Orange rectangles (ORF30) schematize ORF sequences starting with a methionine (green rectangle) and whose length is at least 30bp. The star indicates one peptide found by immunopeptidomics without detecting an ORF.

Supplementary 6

A

TE family proportions within subsets and classes



B

TE family proportions within subsets

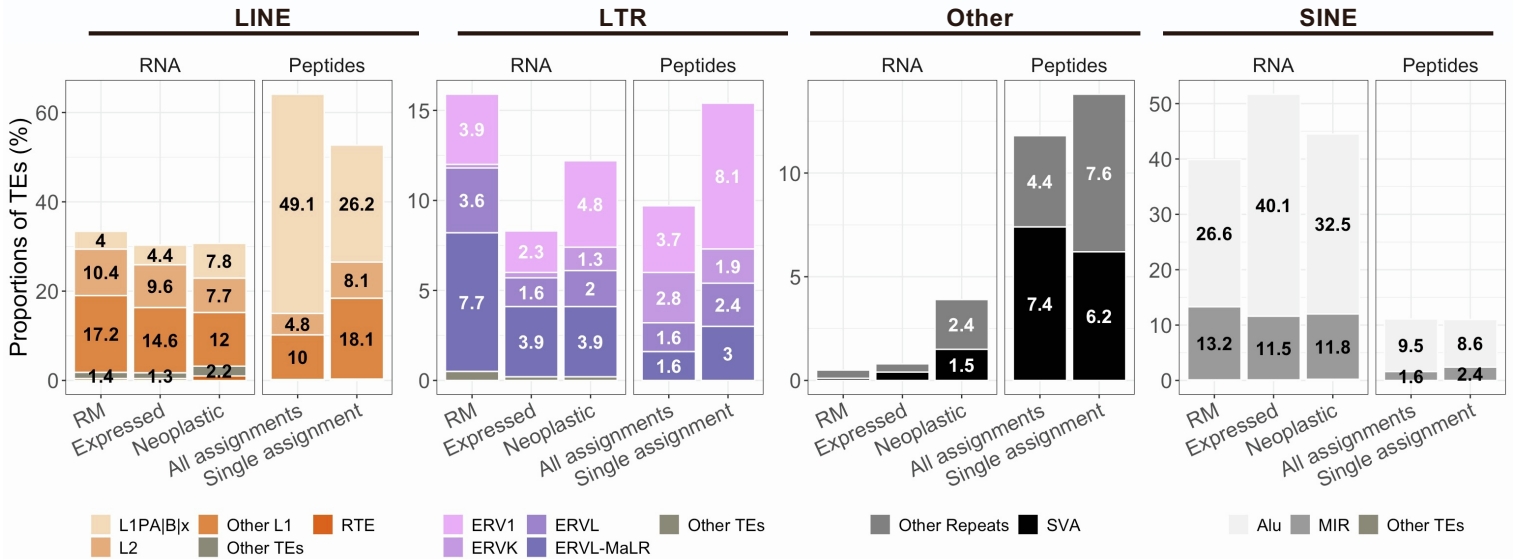
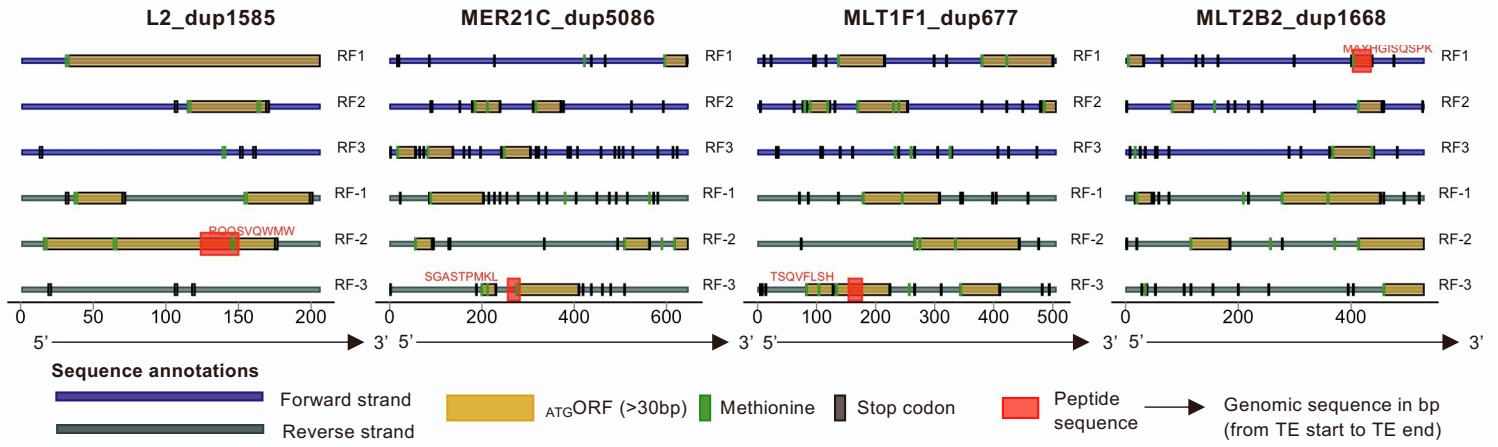


Figure S6. TE family proportions analysis at RNA and peptides levels, Related to Figure 5. (A-B) Barplots representing proportions of TE families by class (A) or globally (B) of all peptide-coding TEs identified with all or single assignment as compared to family proportions of annotated TEs in RepeatMasker, expressed TEs and TEs differentially expressed in neoplastic cells.

Supplementary 7

A Genomic location of TE-derived peptides



B scRNA seq TPM expression in tumor samples

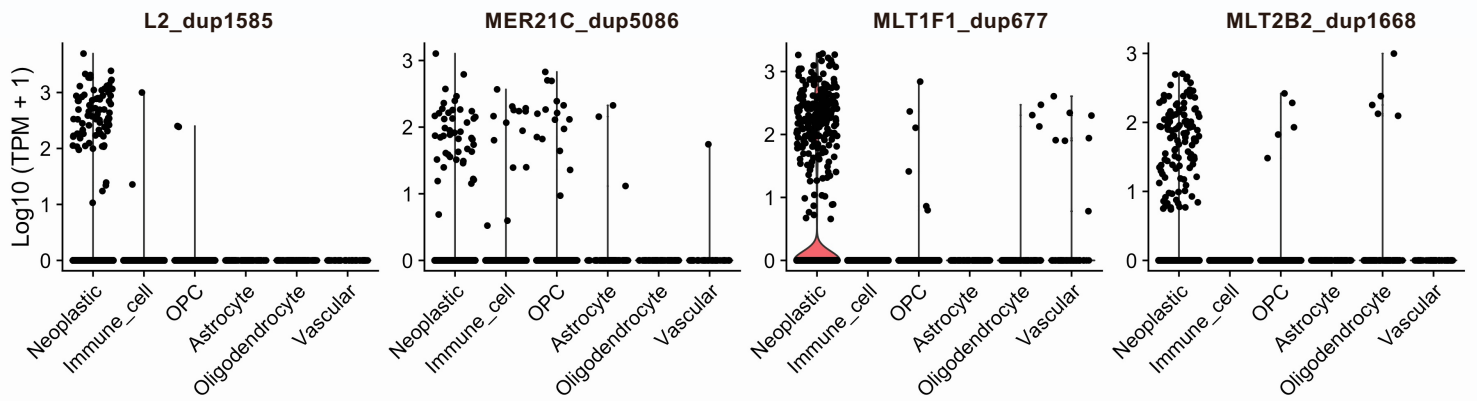
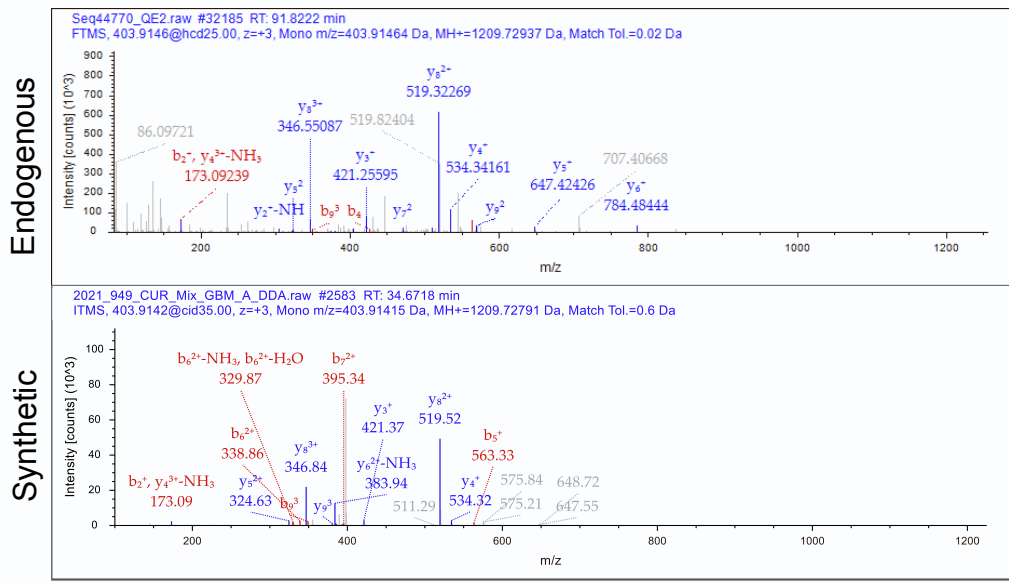


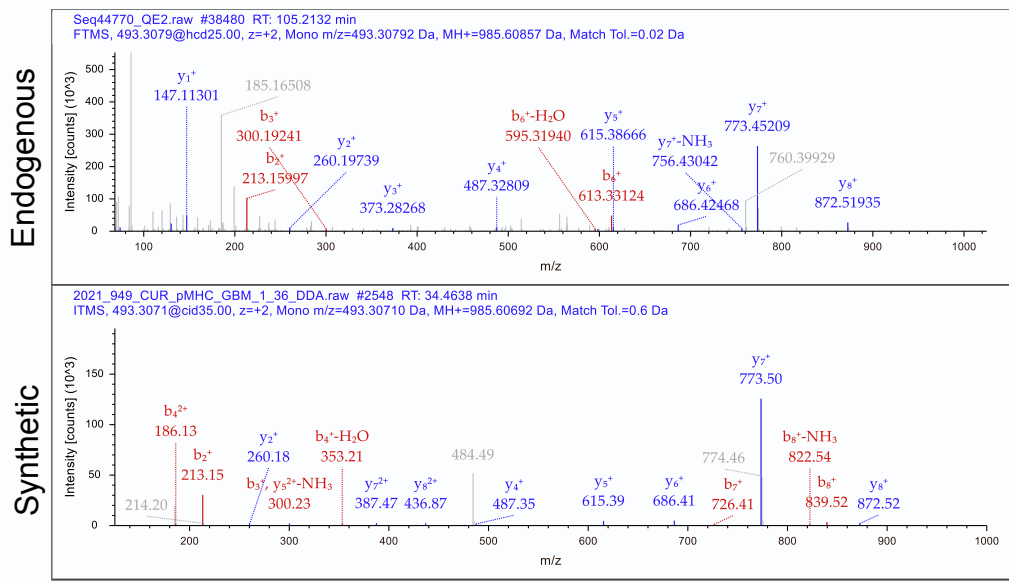
Figure S7. Examples of four TE as tumor-enriched antigens, Related to Figure 6. (A) Graphical representation showing genomic location for four examples marked with a star in (A). All RFs (sense RF1,2,3) and antisense RF-1,-2,-3) are represented. Start codons, stop codons and identified peptides are indicated in green, black and red rectangles, respectively. Orange rectangles (ORF30) schematize ORF sequences starting with a methionine (green rectangle) and whose length is at least 30bp. (B) Violin plots showing the expression in $\log_{10}(\text{TPM}+1)$ for four peptide-coding TEs at the single cell level.

Supplementary Data S2

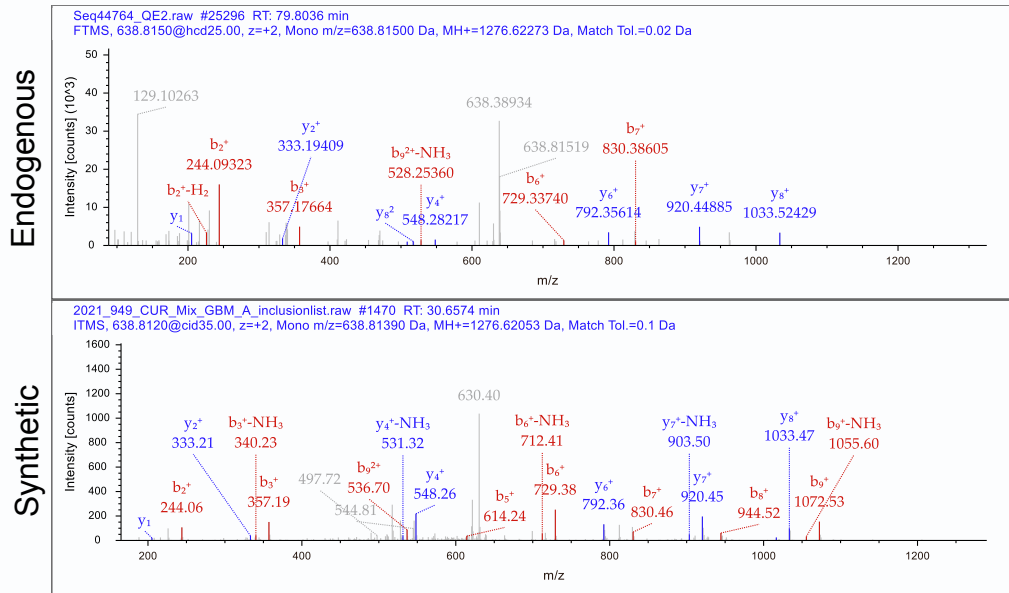
ATPRHLIVRF



IVSAQNILK



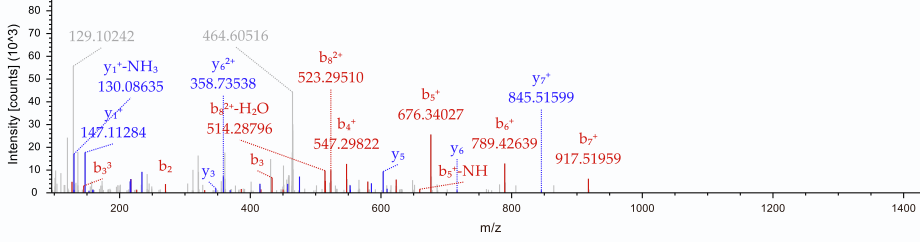
NEIKEDTNKW



RIYNELKQISK

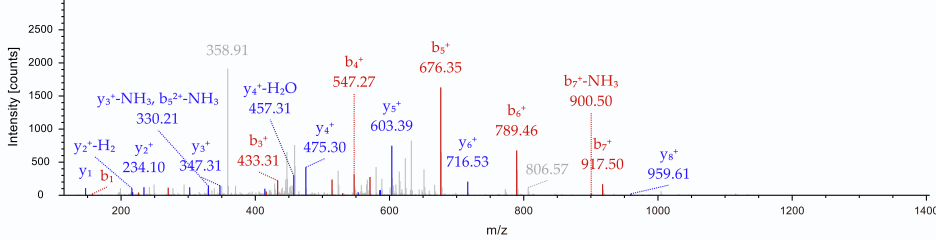
Seq44778_QE2.raw #26948 RT: 84.3052 min
FTMS, 464.6065@hcd25.00, z=+3, Mono m/z=464.6065 Da, MH+=1391.80497 Da, Match Tol.=0.02 Da

Endogenous



2021_949_CUR_Mix_GBM_A_DDA.raw #2243 RT: 32.5699 min
ITMS, 464.6057@cid35.00, z=+3, Mono m/z=464.60565 Da, MH+=1391.80240 Da, Match Tol.=0.1 Da

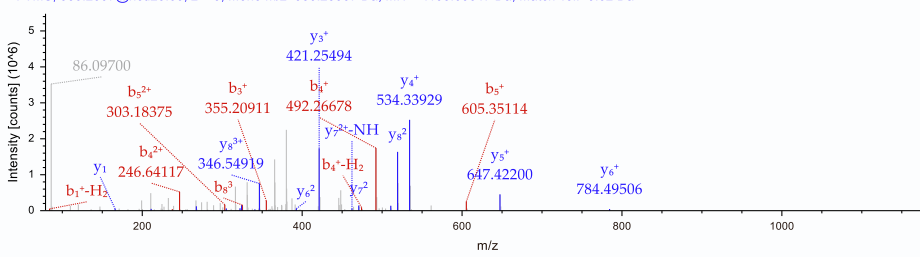
Synthetic



TPRHIVRF

Seq43471_QE2.raw #30418 RT: 85.4976 min
FTMS, 380.2337@hcd25.00, z=+3, Mono m/z=380.23367 Da, MH+=1138.68647 Da, Match Tol.=0.02 Da

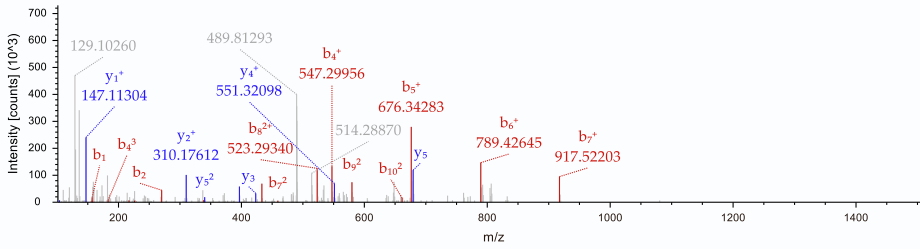
Endogenous



RIYNELKQIYK

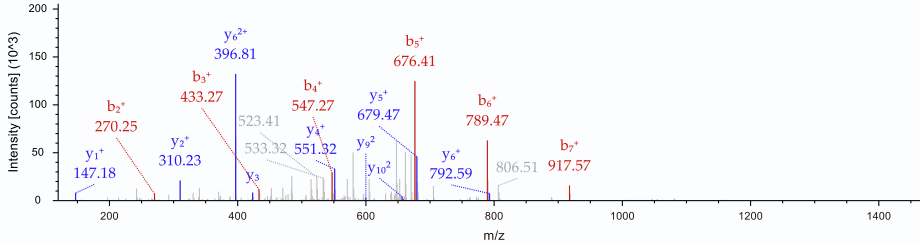
Seq44779_QE2.raw #32870 RT: 94.4966 min
FTMS, 489.9507@hcd25.00, z=+3, Mono m/z=489.95074 Da, MH+=1467.83768 Da, Match Tol.=0.02 Da

Endogenous



2021_949_CUR_Mix_GBM_A_DDA.raw #2189 RT: 32.2532 min
ITMS, 489.9499@cid35.00, z=+3, Mono m/z=489.94986 Da, MH+=1467.83503 Da, Match Tol.=0.1 Da

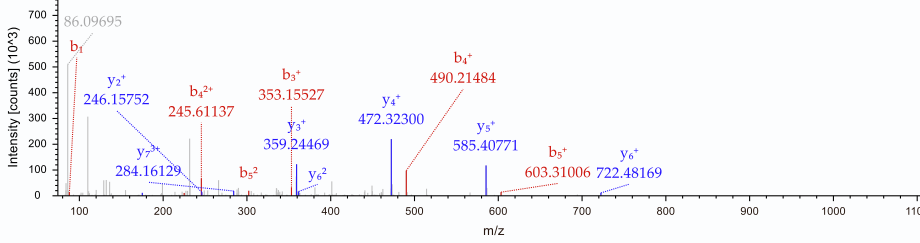
Synthetic



SHQHLLIAR

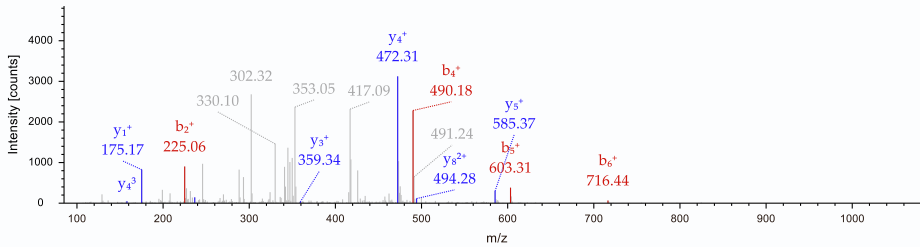
20180504_QEh1_LC1_OC_JMI_HLAIp_HROG02_R2.raw #14137 RT: 31.8507 min
FTMS, 358.8765@hcd27.00, z=+3, Mono m/z=358.87650 Da, MH+=1074.61493 Da, Match Tol.=0.02 Da

Endogenous



2021_949_CUR_Mix_GBM_A_inclusionlist.raw #1436 RT: 30.3391 min
ITMS, 358.8770@cid35.00, z=+3, Mono m/z=358.87836 Da, MH+=1074.62052 Da, Match Tol.=0.1 Da

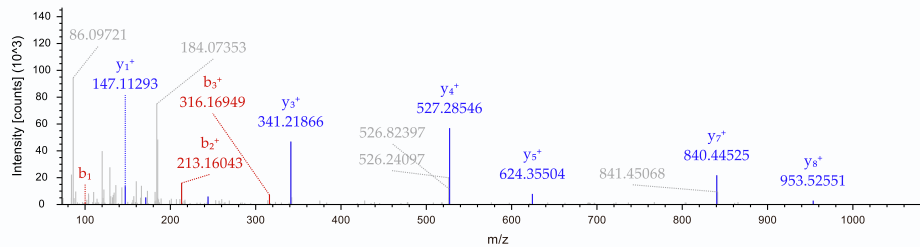
Synthetic



VICLPWPPK

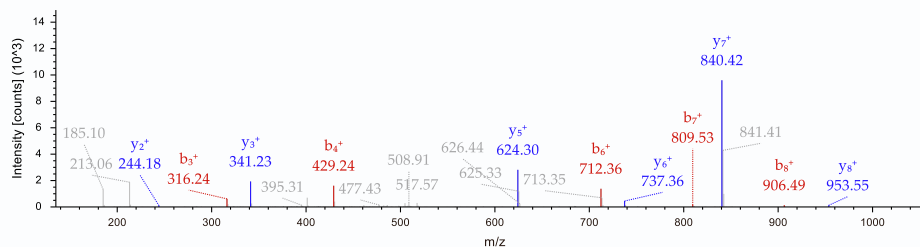
Seq44770_QE2.raw #57889 RT: 148.7909 min
FTMS, 526.8055@hcd25.00, z=+2, Mono m/z=526.80554 Da, MH+=1052.60381 Da, Match Tol.=0.02 Da

Endogenous



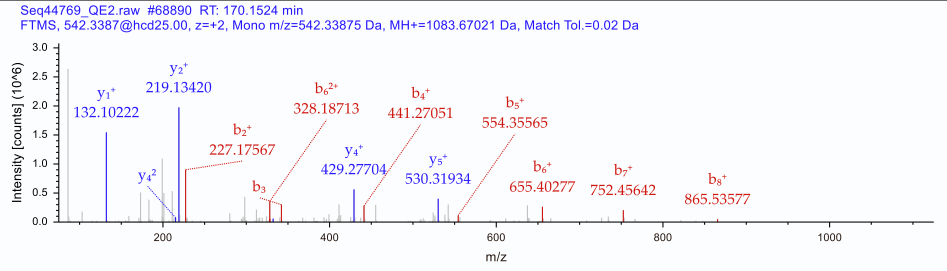
2021_949_CUR_pMHC_GBM_1_36_DDA.raw #3559 RT: 42.0869 min
ITMS, 526.8016@cid35.00, z=+2, Mono m/z=526.80164 Da, MH+=1052.59599 Da, Match Tol.=0.1 Da

Synthetic

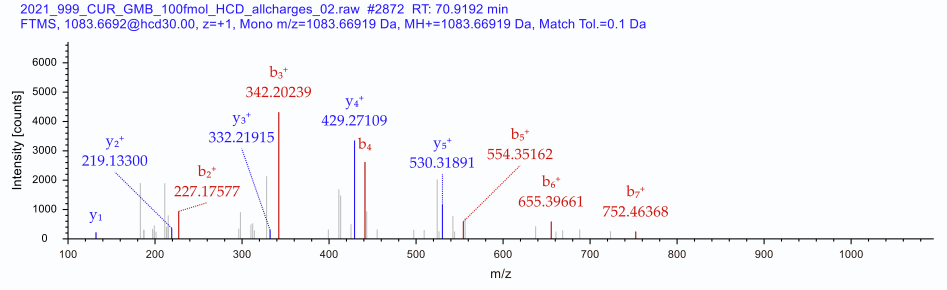


ILDVLTPLSL

Endogenous

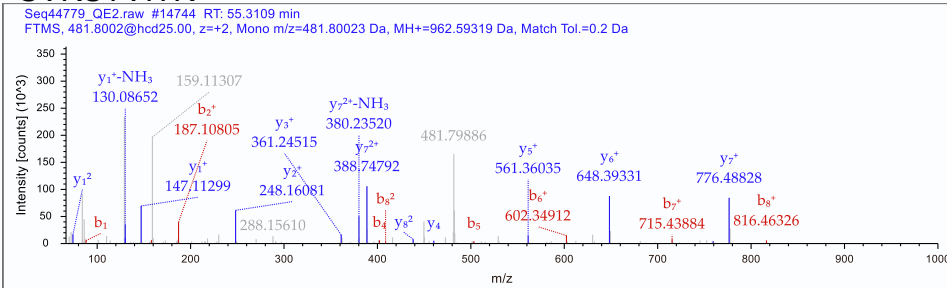


Synthetic

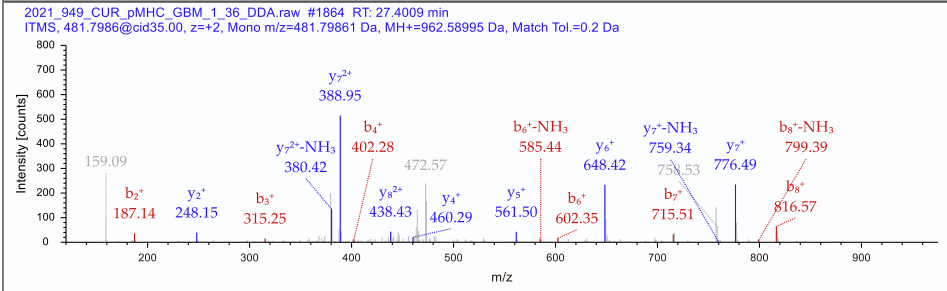


SVKSTVITK

Endogenous

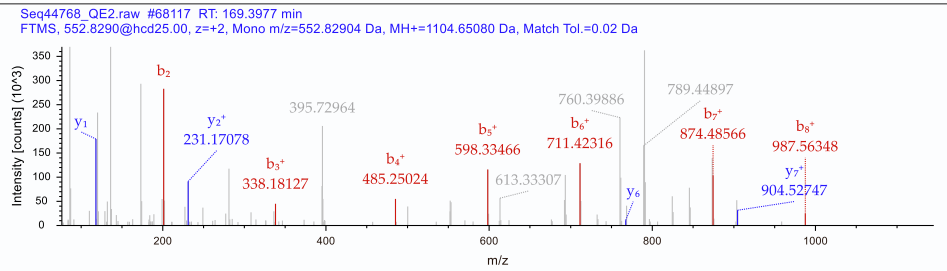


Synthetic

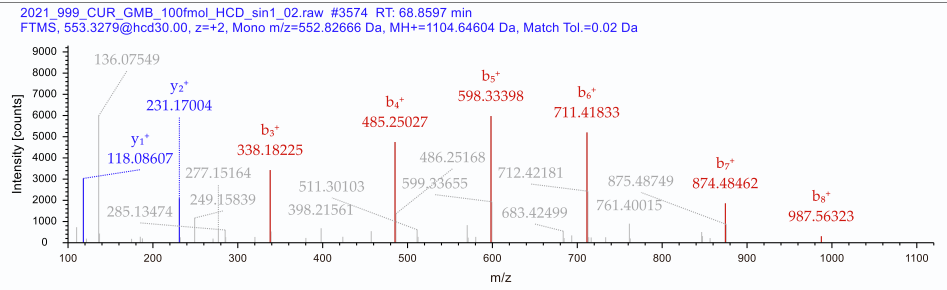


SLHFIYLV

Endogenous

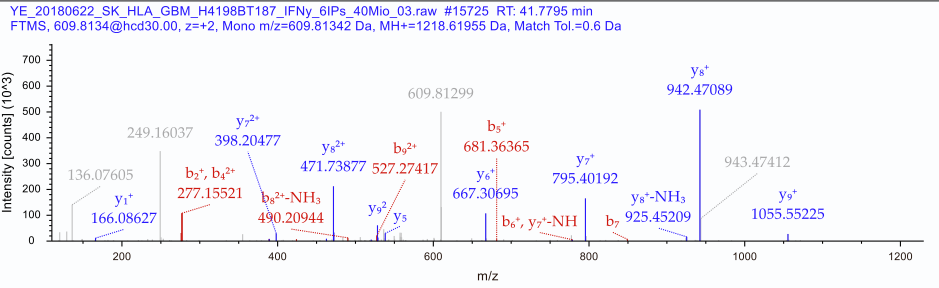


Synthetic

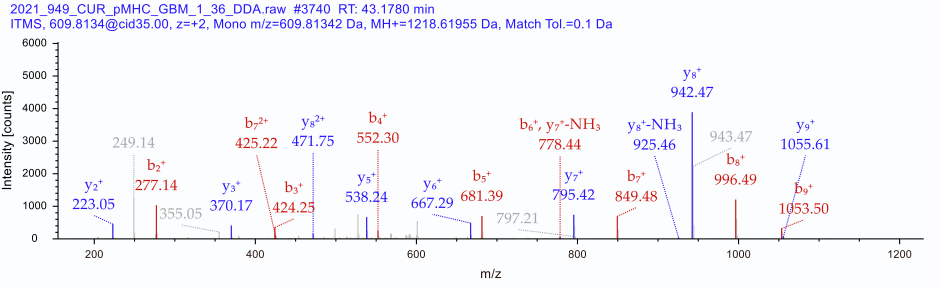


YLFKEPAFGF

Endogenous

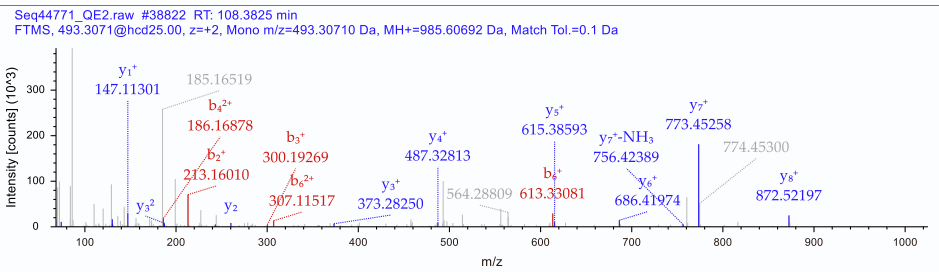


Synthetic

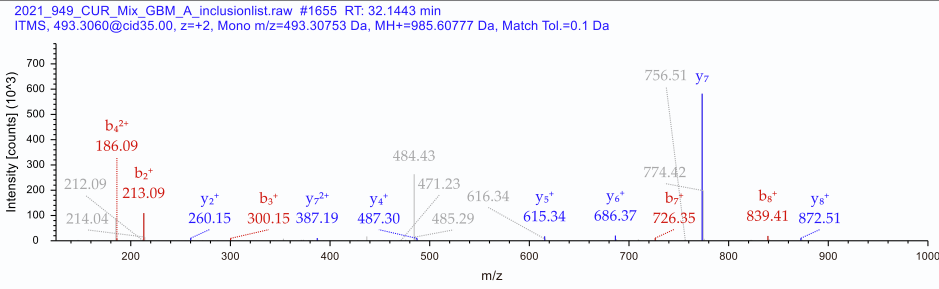


IVSAQNLLK

Endogenous

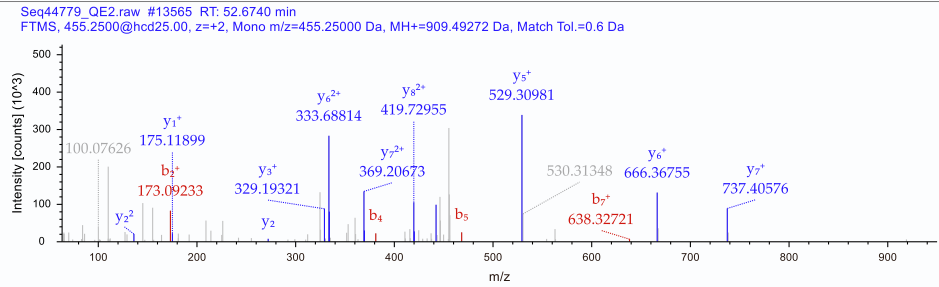


Synthetic

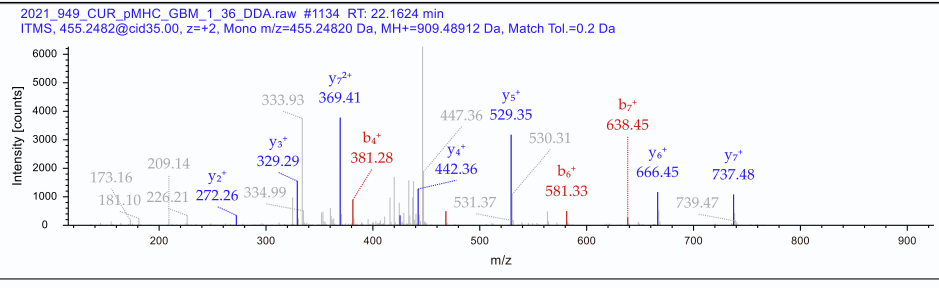


ATAHSLGPR

Endogenous

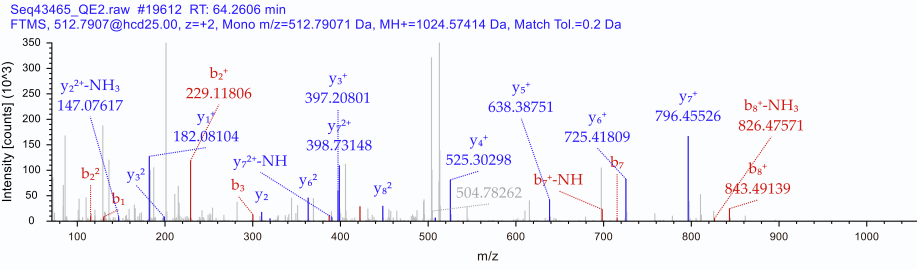


Synthetic

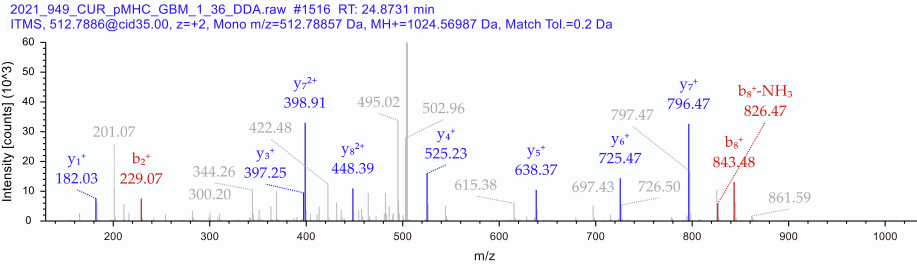


EVASIKSKY

Endogenous

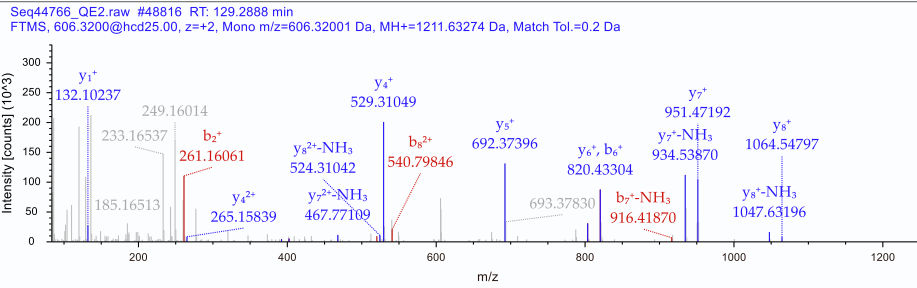


Synthetic

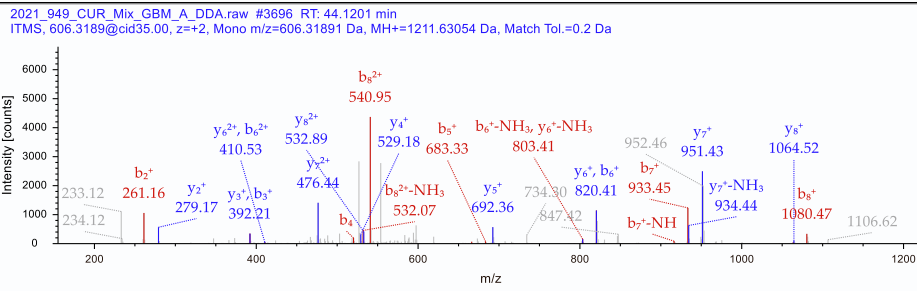


FLMQYHIFL

Endogenous

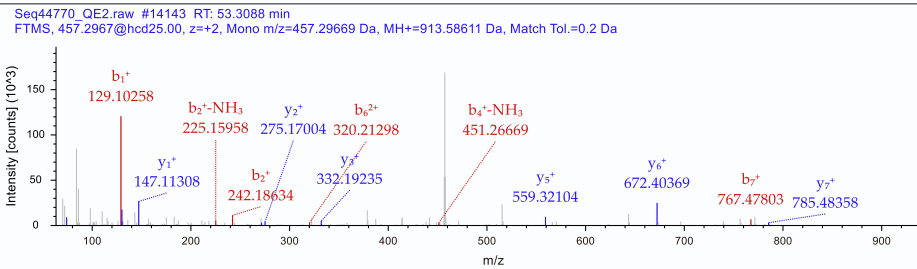


Synthetic

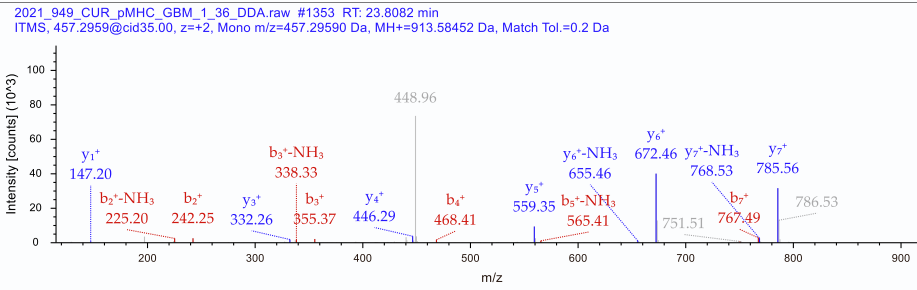


KIILNGQK

Endogenous



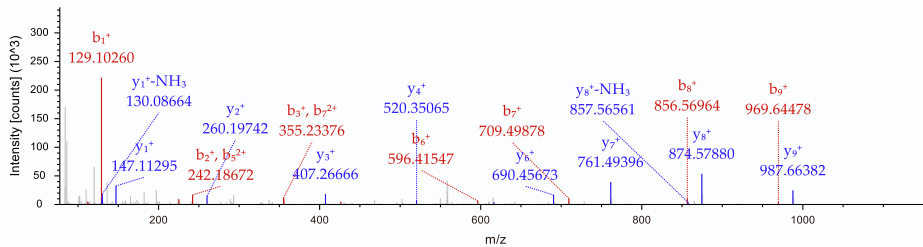
Synthetic



KLIAGLIFLK

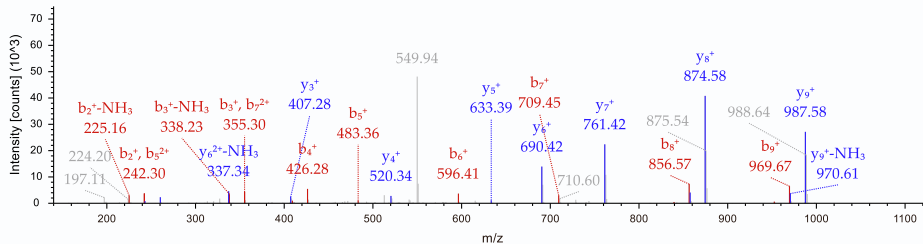
Seq44772_QE2.raw #54884 RT: 142.6314 min
FTMS, 558.3815@hcd25.00, z=+2, Mono m/z=558.38147 Da, MH+=1115.75566 Da, Match Tol.=0.2 Da

Endogenous



2021_949_CUR_pMHC_GBM_1_36_DDA.raw #3639 RT: 42.5866 min
ITMS, 558.3823@cid35.00, z=+2, Mono m/z=558.38226 Da, MH+=1115.75725 Da, Match Tol.=0.2 Da

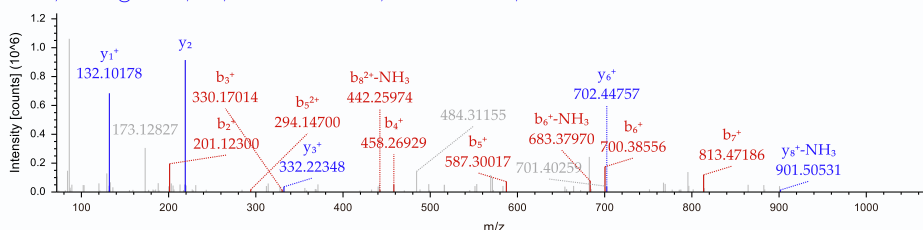
Synthetic



LSEKELISL

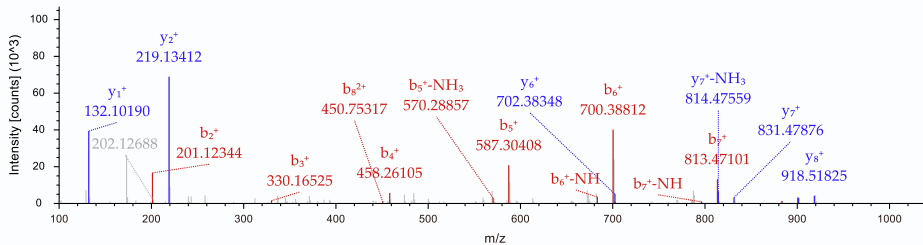
Seq43474_QE2.raw #51075 RT: 130.3443 min
FTMS, 516.3035@hcd25.00, z=+2, Mono m/z=516.30322 Da, MH+=1031.59917 Da, Match Tol.=0.02 Da

Endogenous



2021_999_CUR_GMB_100fmol_HCD_sin1_02.raw #2367 RT: 52.5065 min
FTMS, 517.3058@hcd30.00, z=+2, Mono m/z=516.30481 Da, MH+=1031.60234 Da, Match Tol.=0.2 Da

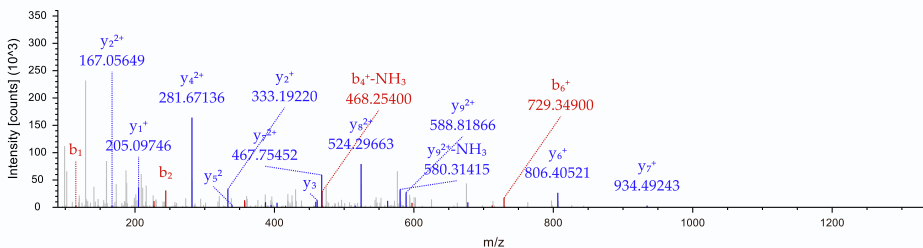
Synthetic



NEIKEDTKKW

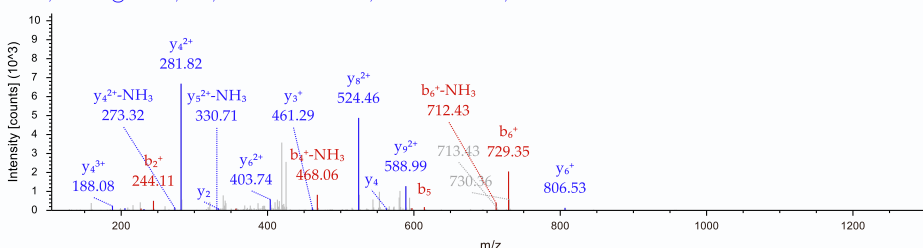
Seq44765_QE2.raw #20277 RT: 68.5736 min
FTMS, 430.8992@hcd25.00, z=+3, Mono m/z=430.89923 Da, MH+=1290.68314 Da, Match Tol.=0.2 Da

Endogenous



2021_949_CUR_pMHC_GBM_1_36_DDA.raw #1623 RT: 25.5514 min
ITMS, 430.8955@cid35.00, z=+3, Mono m/z=430.89551 Da, MH+=1290.67197 Da, Match Tol.=0.2 Da

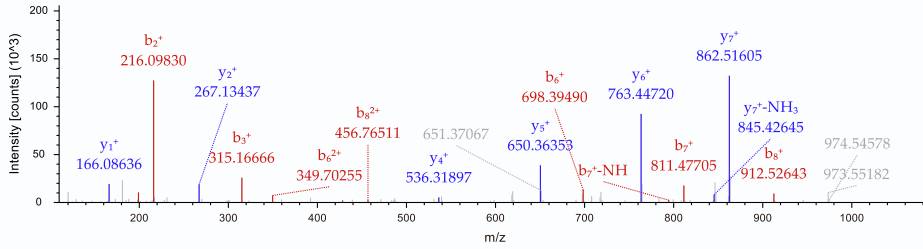
Synthetic



NTVLNRLTF

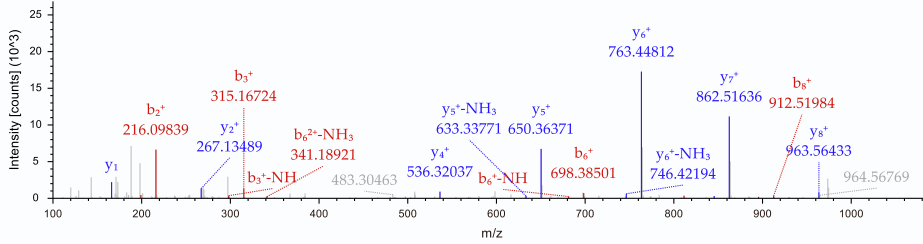
YE_20180622_SK_HLA_GBM_H4198BT187_IFNy_6IPs_40Mo_02.raw #22623 RT: 57.0112 min
 FTMS, 539.3070@hcd30.00, z=+2, Mono m/z=539.30701 Da, MH+=1077.60674 Da, Match Tol.=0.2 Da

Endogenous



2021_999_CUR_GMB_100fmoL_HCD_sin1_02.raw #2402 RT: 52.8911 min
 FTMS, 539.8070@hcd30.00, z=+2, Mono m/z=539.30548 Da, MH+=1077.60369 Da, Match Tol.=0.2 Da

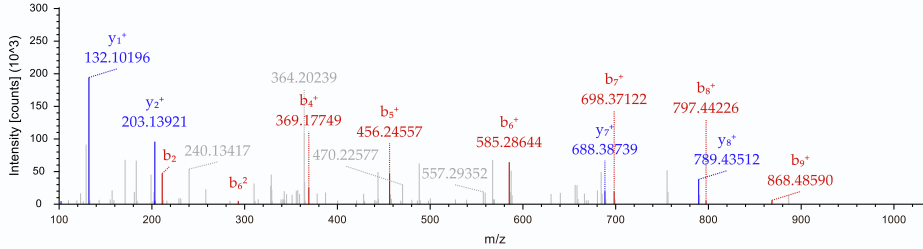
Synthetic



PITGSEIVAI

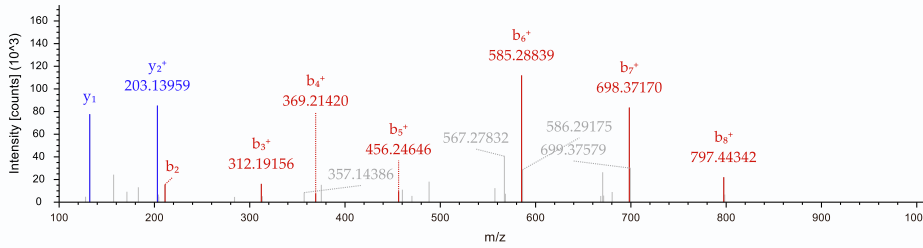
FE20170912_SK_HLA_GBM7_INFy_02.raw #18617 RT: 42.1482 min
 FTMS, 500.2899@hcd25.00, z=+2, Mono m/z=500.28992 Da, MH+=999.57256 Da, Match Tol.=0.2 Da

Endogenous



2021_999_CUR_GMB_100fmoL_HCD_sin1_02.raw #2520 RT: 54.0936 min
 FTMS, 500.7915@hcd30.00, z=+2, Mono m/z=500.29092 Da, MH+=999.57457 Da, Match Tol.=0.2 Da

Synthetic



Data S2 : Spectrums from TE-derived peptides identified in glioblastoma samples, Related to Figure 3. Tumor samples (top) and their comparison with synthetic peptides (bottom) are represented. Spectrums have been extracted from Proteome Discoverer. A total of 23 TE-derived peptides have been validated after comparing the fragmentation patterns between endogenous and synthetic peptides.