Science Translational Medicine
AAAS

www.sciencetranslationalmedicine.org/cgi/content/full/10/470/eaau5516/DC1

Supplementary Materials for

Noncoding regions are the main source of targetable tumor-specific antigens

Céline M. Laumont, Krystel Vincent, Leslie Hesnard, Éric Audemard, Éric Bonneil, Jean-Philippe Laverdure, Patrick Gendron, Mathieu Courcelles, Marie-Pierre Hardy, Caroline Côté, Chantal Durette, Charles St-Pierre, Mohamed Benhammadi, Joël Lanoix, Suzanne Vobecky, Elie Haddad, Sébastien Lemieux, Pierre Thibault, Claude Perreault*

*Corresponding author. Email: claude.perreault@umontreal.ca

Published 5 December 2018, *Sci. Transl. Med.* **10**, eaau5516 (2018)
DOI: 10.1126/scitranslmed.aau5516

The PDF file includes:

Materials and Methods
Fig. S1. Gating strategies for cells isolated by fluorescence-activated cell sorting.
Fig. S2. Architecture of the codes used for our k-mer profiling workflow.
Fig. S3. TSA validation process.
Fig. S4. MS validation of CT26 and EL4 TSA candidates using synthetic analogs.
Fig. S5. Detection of antigen-specific $CD8^+$ T cells in naïve and immunized mice.
Fig. S6. Frequencies of antigen-specific T cells.
Fig. S7. Correlation between antigen-specific T cell frequencies in naïve and immunized mice.
Fig. S8. Purity of the 10H080 B-ALL sample after expansion in NSG mice.
Fig. S9. Overview of the human TEC and mTEC transcriptomic landscapes.
Fig. S10. MS validation of B-ALL TSA candidates using synthetic analogs.
Fig. S11. MS validation of lung cancer TSA candidates using synthetic analogs.
References (*60*–*68*)

Other Supplementary Material for this manuscript includes the following:

(available at www.sciencetranslationalmedicine.org/cgi/content/full/10/470/eaau5516/DC1)

Table S1. Statistics related to the generation of the global cancer databases.
Table S2. Information about samples used in this study.
Table S3. List of CT26 MHC class I–associated peptides.
Table S4. List of EL4 MHC class I–associated peptides.
Table S5. Accession numbers of the ENCODE datasets used in this study.
Table S6. Features of murine TSAs.
Table S7. Experimental values obtained in analyses of mouse TSA immunogenicity.
Table S8. List of 07H103 MHC class I–associated peptides.

Table S9. List of 10H080 MHC class I–associated peptides obtained by mild acid elution.
Table S10. List of 10H080 MHC class I–associated peptides obtained by immunoprecipitation.
Table S11. List of 10H118 MHC class I–associated peptides.
Table S12. List of 12H018 MHC class I–associated peptides.
Table S13. List of lc2 MHC class I–associated peptides.
Table S14. List of lc4 MHC class I–associated peptides.
Table S15. List of lc6 MHC class I–associated peptides.
Table S16. Accession numbers of the Genotype-Tissue Expression (GTEx) datasets used in this study.
Table S17. Features of human TSAs detected in B-ALL specimens.
Table S18. Features of human TSAs detected in lung tumor biopsies.

Materials and Methods

Cell lines. The EL4 T-lymphoblastic lymphoma cell line, the CT26 colorectal cancer cell line and the B-cell hybridoma HB-124 were obtained from the American Type Culture Collection (ATCC). EL4 and CT26 cells were cultured in RPMI 1640/HEPES supplemented with 10% heat-inactivated fetal bovine serum, 1% L-glutamine and 1% penicillin-streptomycin. Cell culture media were further supplemented with 1% non-essential amino acids and 1% sodium-pyruvate or 1% sodium-pyruvate only for EL4 and CT26 cells, respectively. To produce the anti-CDR2 antibody, HB-124 cells were cultured in IMDM supplemented with 10% heat-inactivated fetal bovine serum. Unless stated otherwise, all reagents were purchased from Gibco.

Human primary samples. Primary leukemic samples were expanded in vivo after transplantation in NSG mice as previously described (*59*). Briefly, 1-2 x 10^6 B-ALL cells were thawed and transplanted via i.v. injection into 8–12 week-old sub-lethally irradiated (250 cGy, 137 Cs-gamma source) NSG mice. Mice were sacrificed at signs of disease and cell suspensions were prepared from mechanically disrupted spleens or, for 07H103, from a mix of splenocytes, bone marrow and peritoneal ascites. From there, Ficoll gradients were used to enrich for B-ALL cells prior to isolation of MHC peptides (see section **Isolation of MHC peptides**). After Ficoll gradient, purity and viability of each B-ALL sample were assessed using flow cytometry and one representative experiment is shown in **fig. S8**. Briefly, 0.5 x 10^6 cells were stained with Pacific Blue anti-human CD45 (BioLegend), PE-Cy7 anti-human CD19 (BD Bioscience), APC-eFluor780 anti-mouse CD45.1 (eBioscience) and 7-AAD (BD Bioscience,). B-ALL cells were defined as 7-AAD-huCD45$^+$huCD19$^+$. Data acquisition was performed on a BD Canto II cytometer (BD Bioscience). The analysis was done with BD FACSDiva 4.1 software. For all

samples, HLA typing was obtained using OptiType version 1.0, running with default parameters for RNA-sequencing (RNA-Seq) data (see section **RNA extraction, library preparation and sequencing**).

Peptides. Native and ^{13}C-labelled versions of tumor-specific antigens (TSAs) were synthesized by GenScript. For the ^{13}C-analogs, the labelled amino acids are underlined in the following list: VNYIHRNV, VNYLHRNV, TVPLNHNTL, VTPVYQHL, IILEFHSL. Purity, as determined by the manufacturer, was greater than 95% and 75% for native and ^{13}C-labelled peptides, respectively.

Murine mTEC^{hi} extraction. Thymi were isolated from 5–8 week-old C57BL/6 or Balb/c mice and mechanically disrupted to extract thymocytes. Stromal cell enrichment was performed as previously described (*60*). Thymic stromal cells were stained with biotinylated Ulex europaeus lectin 1 (UEA1; Vector Laboratories), PE-Cy7–conjugated streptavidin (BD Biosciences), and the following antibodies: Alexa Fluor 700 or FITC anti-CD45, PE anti–I-A^b (BD Biosciences), Alexa Fluor 700 I-A/I-E, APC-Cy7 anti-EpCAM (BioLegend). Cell viability was assessed using 7-aminoactinomycin D (7-AAD; BD Biosciences). Live mature medullary thymic epithelial cells (mTEC^{hi}) were gated as 7-AAD$^-$ CD45$^-$ EpCAM$^+$ UEA1$^+$ MHC II^{hi} and sorted on a FACS AriaIIIu (BD Biosciences, **fig. S1A**).

Human TEC and mTEC extraction. Thymi obtained from 3-month-old to 7-year-old individuals were kept at 4ºC in 50 ml conical tubes containing media and cut in 2-5 mm cubes within hours following their surgical resection. For long-term preservation, thymic cubes were

frozen in cryovials containing heat-inactivated human serum / 10% DMSO and kept in liquid nitrogen for a maximum of 3 years. Cryopreserved thymic samples were transferred to our laboratory on dry ice and used to isolate human TEC and mTEC following a protocol adapted from C. Stoeckle *et al.* (*61*). Thymic tissue was cut into small fragments, then digested at 37°C using a solution of 2 mg/mL Collagenase A (Roche) and 0.1 mg DNase I/ml (Sigma-Aldrich) in RPMI-1640 (Gibco) for three to five periods of 40 min. After the second digestion, a solution of Trypsin/EDTA (Gibco) was added, for which the activity was neutralized by adding FBS (Invitrogen) 15 min before the end of incubation. For TEC and mTEC sorting (**fig. S1C**), cell suspensions were stained with Pacific blue-conjugated anti-CD45 (BioLegend), PE-conjugated anti-HLA-DR (BioLegend), APC-conjugated anti-EpCAM (BioLegend), Alexa 488-conjugated anti-CDR2 (produced in our lab with the HB-124 hybridoma – see section **Cell lines** – and conjugated with the Dylight 488 Fast conjugation kit from abcam, only for mTEC samples) and cell viability was assessed using 7-AAD (BD Biosciences).

RNA extraction, library preparation and sequencing. For EL4 and CT26 cells, one replicate of 5 x 10^6 cells was used to perform RNA-Seq. For C57BL/6 and Balb/c mTEC^{hi}, RNA-Seq was performed in triplicate on a minimum of 31,686 or 16,338 FACS-sorted cells extracted from 2 females and 2 males. For primary leukemic cells, RNA-Seq was performed on a single replicate of 2.0 to 4.0 x 10^6 cells. For human TEC and mTEC, we performed one RNA-Seq replicate per donor with 33,076 to 84,198 FACS-sorted TECs or 50,058 to 100,719 mTECs. In all cases, total RNA was isolated using TRIzol (Invitrogen), further purified using the RNeasy kit or RNeasy micro kit (Qiagen) as recommended by each manufacturer. For each lung tumor biopsy (three in total), total RNA was isolated from ~30 mg of tissues using the AllPrep DNA/RNA/miRNA

Universal kit (Qiagen) as recommended by the manufacturer and was used to perform one replicate of RNA-Seq per sample. Each murine sample (EL4, CT26 and murine mTEC^{hi}) were quantified on a Nanodrop 2000 (Thermo Fisher Scientific) and RNA quality was assessed on a 2100 Bioanalyzer (Agilent Genomics) in order to select samples with an RNA integrity number ≥ 9. For human samples (B-ALLs, lung tumor biopsies and human TEC/mTEC), quantification of total RNA was made by QuBit (ABI) and quality of total RNA was assessed with the 2100 BioAnalyzer (Agilent Genomics) in order to select samples with an RNA integrity number ≥ 7. cDNA libraries were prepared from 2-4 µg for EL4 and CT26 cells, 50-100 ng for murine mTEC^{hi}, 500 ng for B-ALLs specimens, 4 µg for lung tumor biopsies, 8-13 ng for human TECs or 41-68 ng for human mTECs of total RNA using the TruSeq Stranded Total RNA Library Prep Kit (EL4 cells), KAPA Stranded mRNA-Seq Kit (CT26 cells, C57BL/6 mTEC^{hi}, human mTEC, lung tumors and B-ALL specimens) or KAPA RNA HyperPrep Kit (Balb/c mTEC^{hi}, human TEC). These libraries were further amplified by 9-16 cycles of PCR before sequencing. Paired-end RNA-Seq was performed on an Illumina NextSeq 500 (Balb/c mTEC^{hi}, human TEC and mTEC) or HiSeq 2000 (any other sample) and yielded an average of 175 and 199 x 10^6 reads per murine and human sample, respectively.

Generation of canonical cancer and normal proteomes. For all samples, RNA-Seq reads were trimmed for sequencing adapters and low quality 3' bases using Trimmomatic version 0.35 and then aligned to the reference genome, GRCm38.87 for murine samples and GRCh38.88 for human samples, using STAR version 2.5.1b (*62*) running with default parameters except for --alignSJoverhangMin, --alignMatesGapMax, --alignIntronMax, and --alignSJstitchMismatchNmax parameters for which default values were replaced by 10, 200,000,

200,000 and "5 -1 5 5", respectively. Single-base mutations with a minimum alternate count setting of 5 were identified using freeBayes version 1.0.2-16-gd466dde (arXiv:1207.3907) and exported in a VCF, which was converted to an agnostic single-nucleotide polymorphism file format compatible with pyGeno (*63*). Finally, transcript expression was quantified in transcripts per million (tpm) with kallisto version 0.43.0 (https://pachterlab.github.io/kallisto/about) running with default parameters. Of note, kallisto index was constructed using the index functionality and using the appropriate *.cdna.all.fa.gz files downloaded from Ensembl (*64*). To build each sample's canonical proteome, we used pyGeno to (i) insert high-quality sample-specific single-base mutations (freeBayes quality > 20) in the reference exome, thereby creating a personalized exome, and to (ii) export sample-specific sequence(s) of known proteins generated by expressed transcripts (tpm > 0). These protein sequences were written to a fasta file that was subsequently used for mass spectrometry (MS) database searches (Cancer canonical proteome) and/or MHC peptide classification (Cancer and normal canonical proteome). See **Fig. 1A** for a schematic and **table S1** for statistics.

Generation of cancer and normal k-mer databases. For all cancer and normal samples, both R1 and R2 fastq files were independently downloaded and trimmed for sequencing adapters and low quality 5' and 3' bases using Trimmomatic version 0.35. To ensure that all reads were on the transcript-encoding strand, R1 reads were reverse complemented using the **fastx_reverse_complement** function of the FASTX-Toolkit version 0.0.14. Using Jellyfish version 2.2.3 (*34*), we then generated 33- and 24-nucleotide-long k-mer databases, required for k-mer profiling and MHC peptide classification, respectively. See **fig. S2A** for details. Of note, when multiple biological replicates (murine mTEC^{hi}) or when multiple samples from unrelated

donors (human TEC and mTEC) were available, fastq files were concatenated to generate a single normal k-mer database per condition (C57BL/6, Balb/c or human).

k-mer filtering and generation of cancer-specific proteomes. To extract 33-nucleotide-long k-mers that could give rise to TSAs, we applied a sample-specific threshold on k-mer occurrence in order to exclude sequencing errors: at least 4 times in EL4 or CT26 cells, 7 times in lung tumor biopsies and 10 times in primary leukemic samples. Cancer-specific k-mers were then obtained by excluding those expressed in the relevant murine $mTEC^{hi}$ or human TEC/mTEC k-mer database (see **fig. S2B**). This cancer-specific k-mer set was further assembled into longer sequences, called contigs. Briefly, one of the submitted 33-nucleotide-long k-mer is randomly selected to be used as a seed that is then extended from both ends with consecutive k-mers overlapping by 32 nucleotides on the same strand (-r option disabled, as we were working with stranded sets of k-mers). The assembly process stops when either no k-mers can be assembled, meaning that no 32-nucleotide-overlapping k-mer can be found, or when more than one k-mer fits (–a 1 option for linear assembly). If so, a new seed is selected and the assembly process resumes until all k-mers from the submitted list have been used once (see **fig. S2C**). This step is done by the **kmer_assembly** tool from NEKTAR, an in-house developed software. To obtain amino acid sequences, we 3-frame translated contigs that were at least 34 nucleotides long using an in-house python script. Cancer-specific amino acid sequences were then split at internal stop codons and resulting subsequences of at least 8 amino acids long were given a unique ID before being included in the relevant cancer-specific proteome (see **fig. S2D**). See **Fig. 1B** for schematic and **table S1** for statistics.

Isolation of MHC peptides. For EL4 and CT26 cells, three biological replicates of 250 x 10^6 cells were prepared from exponentially growing cells. For all primary leukemic samples, three biological replicates of ~450 to 700 x 10^6 cells were prepared from freshly harvested leukemic cells (see section **Human primary samples**). MHC peptides were obtained as previously described (*65*), with minor modifications: following mild acid elution (MAE), peptides were desalted on an Oasis HLB cartridge (30 mg, Waters) and filtered on a 3 kDa molecular weight cut-off (Amicon Ultra-4, Millipore) to remove β2m proteins. For one of our primary leukemic samples (specimen 10H080), we prepared four additional replicates of 100 x 10^6 cells and isolated MHC peptides by immunoprecipitation (IP) as previously described (*59*). Finally, lung tumor biopsies (wet weight ranging from 771 to 1,825 mg, see section **Study design**) were cut in small pieces (cubes of ~3 mm in size) and 5 ml of ice-cold PBS containing protein inhibitor cocktail (Sigma) was added to each tissue sample. Tissues were first homogenized twice using an Ultra Turrax T25 homogenizer (20 seconds at 20,000 rpm, IKA-Labortechnik) and then once using an Ultra Turrax T8 homogenizer (20 seconds at 25,000 rpm, IKA-Labortechnik). Then, 550 µl of ice-cold 10X lysis buffer (10% w/v CHAPS) was added to each sample and MHC peptides were immunoprecipitated as previously described (*59*) using 1 mg (1 ml) of covalently cross-linked W6/32 antibody to protein A magnetic beads per sample. Regardless of the isolation technique, peptide extracts were all dried using a Speed-Vac and kept frozen prior to MS analyses.

Mass spectrometry analyses. Dried peptide extracts were re-suspended in 0.2 % formic acid. For EL4 and CT26, peptide extracts were loaded on a home-made C_{18} pre-column (5 mm x 360 μm i.d. packed with C_{18} Jupiter Phenomenex) and separated on a home-made C_{18} analytical

column (15 cm x 150 µm i.d. packed with C_{18} Jupiter Phenomenex) with a 56-min gradient from 0–40 % acetonitrile (0.2 % formic acid) and a 600 $nl.min^{-1}$ flow rate on a nEasy-LC II system. For all human samples, peptide extracts were loaded on a home-made C_{18} analytical column (15 cm x 150 µm i.d. packed with C_{18} Jupiter Phenomenex) with a 56-min gradient from 0–40 % acetonitrile (0.2 % formic acid, 07H103, 10H080-MAE, 10H118 and 12H018) or with a 100-min gradient from 5–28 % acetonitrile (0.2 % formic acid, lung tumor biopsies and 10H080-IP) and a 600 nl.min-1 flow rate on a nEasy-LC II system. Samples were analyzed with a Q-Exactive Plus (EL4, Thermo Fisher Scientific) or HF (all other samples, Thermo Fisher Scientific). For the Q-Exactive Plus, each full MS spectrum, acquired with a 70,000 resolution, was followed by 12 MS/MS spectra, where the most abundant multiply charged ions were selected for MS/MS sequencing with a resolution of 17,500, an automatic gain control target of 1e6, an injection time of 50 ms and a collision energy of 25 %. For the Q-Exactive HF, each full MS spectrum, acquired with a 60,000 resolution, was followed by 20 MS/MS spectra, where the most abundant multiply charged ions were selected for MS/MS sequencing with a resolution of 15,000 (CT26, 07H103, 10H080-MAE, 10H118, 12H018) or 30,000 (lung tumor biopsies, 10H080-IP), an automatic gain control target of 5e4, an injection time of 100 ms and a collision energy of 25 %. Peptides were identified using Peaks 8.5 (Bioinformatics Solution Inc.) and peptide sequences were searched against the relevant global cancer database, obtained by concatenating the canonical cancer proteome and cancer-specific proteome (see sections **Generation of canonical cancer and normal proteomes** and **k-mer filtering and generation of cancer-specific proteomes**). For peptide identification, tolerance was set at 10 ppm and 0.01 Da for precursor and fragment ions, respectively. Occurrence of oxidation (M) and deamidation (NQ) were considered as post-translational modifications.

Identification of MHC peptides. To select for MHC peptides, lists of unique identifications obtained from Peaks were filtered to include 8–11-amino-acid-long peptides that had a percentile rank ≤ 2% for at least one on the relevant MHC I molecules, as predicted by NetMHC 4.0 (*66*). We did not compute false discovery rates on these lists since we wanted to identify as many TSAs as possible. However, we applied a sample-specific threshold on the Peaks score to guaranty that our list of MHC peptides only included 5% of decoy identifications $\left(\%_{decoy} = \left(\#_{decoys} / \#_{targets}\right) \times 100\right)$. MHC peptides identified in each samples are reported in **tables S3, S4** and **S8-S15**.

Identification and validation of TSA candidates. To identify TSA candidates, each MHC peptide and its coding sequence were queried to the relevant cancer and normal canonical proteomes or cancer and normal 24-nucleotide-long k-mer databases, respectively. Here, the normal canonical proteome and normal 24-nucleotide-long k-mer database were built using (i) RNA-Seq data from syngeneic $mTEC^{hi}$ for EL4 and CT26 cells and (ii) RNA-Seq data from two TECs and four mTECs samples for all human tumor samples. MHC peptides detected in the normal canonical proteome were excluded regardless of their coding sequence detection status, as they are likely to be tolerogenic. MHC peptides that were truly cancer-specific, in other words, which were neither detected in the normal canonical proteome nor in normal k-mers, were flagged as TSA candidates. MHC peptides absent from both canonical proteomes but present in both k-mer databases needed to have their RNA coding sequence overexpressed by at

least 10-fold in cancer cells compared to normal cells in order to be flagged as such (see **fig. S2A**). Finally, MHC peptides corresponding to several RNA sequences (derived from different proteins) could only be flagged as TSA candidate if their respective coding sequences were concordant, that is, if those consistently flagged the relevant peptide as a TSA candidate. MS/MS spectra of all TSA candidates were manually inspected to remove any spurious identifications. Besides, sequences presenting with multiple genomically possible I/L variants were further inspected to report both variants when they were distinguishable by MS, or only the most expressed variant when they were not (see **fig. S3B**). Finally, we assigned a genomic location to all those MS-validated TSA candidates by mapping reads containing MHC peptide-coding sequences on the reference genome (GRCm38.87 or GRCh38.88) using BLAT (tool from the UCSC genome browser). TSA candidates for which reads did not match to a concordant genomic location or which matched to hypervariable regions (such as the MHC, Ig or TCR genes) or multiple genes were excluded. For those with a concordant genomic location, we used IGV (*67*) to exclude TSA candidates with a MHC peptide-coding sequence overlapping synonymous mutations with regard to their relevant normal counterpart or, for human TSA candidates, those overlapping a known germline polymorphism (listed in dbSNP v. 149, **fig. S3C**). Remaining peptides were classified as mTSAs or aeTSA candidates, depending if their coding sequence overlapped a cancer-specific mutation or not.

Peripheral expression of MHC peptide-coding sequences. To assess the peripheral expression of tumor-associated antigens' and aeTSA candidates' peptide-coding sequences, we used RNA-Seq data from 22 murine tissues, which had been sequenced by the ENCODE (*39, 40*) consortium (**table S5**) or from 28 peripheral human tissues (~50 donors per tissue), which had

been sequenced by the GTEx consortium and downloaded from the GTEx Portal on 04/16/2018 (phs000424.v7.p2, **table S16**). Briefly, RNA-Seq data from each tissue were transformed into 24-nucleotide-long k-mer databases with Jellyfish 2.2.3 (using the –C option) and used to query each peptide-coding sequence's 24-nucleotide-long k-mer set. For each RNA-Seq experiment, the number of reads fully overlapping a given MHC peptide-coding sequence ($r_{overlap}$) was estimated using the k-mer set's minimum occurrence (k_{min}). Indeed, we hypothesized that k_{min} ~ $r_{overlap}$ because, except for low complexity RNA-Seq reads that might generate the same k-mer multiple times, one k-mer always originate from a single RNA-Seq read. Thus, to compare the peptide-coding sequence expression across all tissues, we transformed this $r_{overlap}$ value into a number of reads detected per 10^8 reads sequenced ($rphm$) using the following formula: $rphm = \frac{(r_{overlap} \times 10^8)}{r_{tot}}$, with r_{tot} representing the total number of reads sequenced in a given RNA-Seq experiment. These values were then log-transformed ($\log_{10}(rphm + 1)$) and averaged across all RNA-Seq experiments of a given tissue. aeTSA candidates exhibiting a peripheral expression in 10 or less tissues (at rphm > 0) or in less than 5 tissues other than the liver (at rphm > 15) for murine and human candidates respectively, were considered as genuine aeTSAs. Features of those aeTSAs, as well as mTSAs are reported in **tables S6, S17 and S18**.

MS validation of TSA candidates. For CT26 TSA candidates and two EL4 TSA candidates (ATQQFQQL and SSPRGSSTL), we compared the previously acquired MS/MS spectra to the relevant ^{12}C-analog. For the other five EL4 TSA candidates tested in vivo (IILEFHSL, TVPLNHNTL, VNYIHRNV, VNYLHRNV, VTPVYQHL), we eluted MHC peptides from six additional EL4 replicates (~450 to 1,400 x 10^6 cells per replicate), which were all processed as previously described (see Sections **Isolation of MHC peptides** and **Mass spectrometry**

analyses). For absolute quantification, three of the six EL4 replicates were spiked with 500 fmol of each ^{13}C-labelled TSA. For sequence validation, MS/MS spectrum of ^{12}C TSA candidates were acquired prior to sample analysis by PRM MS. Briefly, the PRM acquisition, which monitored five peptides as scheduled (each peptide is only monitored in a 10-minute window centered on its elution time), consisted of one MS1 scan followed by the targeted MS/MS scans in HCD mode. Automatic gain controls and injection times for the survey scan and the tandem mass spectra were 3e6 – 50 ms and 2e5 – 100 ms, respectively. In all cases, Skyline (*68*) was used to extract the endogenous MS/MS spectrum of each TSA candidate and compare it to the relevant ^{12}C MS/MS spectrum (sequence validation) or to extract the intensity of the endogenous and the relevant synthetic ^{13}C-labelled peptide (absolute quantification). Using the following formula, these intensities were further used to compute the number of TSA copy per cell for each replicate: $\left(n_{synthetic} \times I_{endogenous} \times N_A/I_{synthetic}\right) \times (1/N_{cells})$ with $n_{synthetic}$, initial number of moles spiked for the considered synthetic ^{13}C-labelled TSA; $I_{endogenous}$ and $I_{synthetic}$, intensity of the relevant endogenous and ^{13}C-labelled TSA, respectively; N_A, Avogadro's number; N_{cells}, initial number of cells used for mild acid elution.

Cumulative number of transcripts detected in human TEC and mTEC samples. Restricting our analysis to transcripts expressed at a tpm > 1 in at least one of our six samples (2 TECs and 6 mTECs), we first computed Spearman's rank correlation coefficient for each 1-to-1 TEC/mTEC comparison. Then, using those same sets of expressed transcripts, we computed the cumulative numbers of transcripts (cT) detected as each additional sample are analyzed. Because the order in which samples are introduced in the analysis can influence cT values, we averaged the cT values across all sample permutations and used those average data points to fit the following predictive

curve (with the R's 'nls' function): $cT = \frac{a \times (nS-1)}{[b+ (nS-1)]} + c$, with cT, the cumulative numbers of transcripts and nS, the number of analyzed samples. This equation was then used to extrapolate the number of transcripts that would have been detected by studying up to 20 samples and which can be estimated by simply computing $\lim_{nS\to\infty}(cT)$.

Generation of bone marrow-derived dendritic cells (DCs), mouse immunization and EL4 cell injection. Bone marrow-derived DCs were generated as previously described *(42)*. For mouse immunization, DCs from male C57BL/6 mice were pulsed with 2 μM of the selected peptide for 3 hours, then washed. 8–12 week-old female C57BL/6 mice were injected i.v. with 10^6 DCs pulsed with one of the TSA or with irradiated EL4 cells (10,000 cGy) at day −14 and −7. As negative control, C57BL/6 female mice were immunized with unpulsed DCs. At day 0 and day 150, mice were injected i.v. with 5 x 10^5 EL4 cells and were monitored for weight loss, paralysis, or tumor outgrowth.

IFN-γ ELISpot and avidity assays. IFN-γ ELISpot and avidity assays were performed as previously described *(42)*. Briefly, Millipore MultiScreen PVDF plates were permeabilized with 35% ethanol, washed, and coated overnight using the Mouse IFN-γ ELISpot Ready-SET-Go! reagent set (eBioscience). At day 0 following mice immunization, splenocytes were harvested from immunized or naive mice. 30 x 10^6 splenocyte/mL were stained with FITC-conjugated anti-CD8a (BD Biosciences) for 30 minutes at 4°C, washed, and sorted using a FACS AriaIIu or a FACS AriaIIIu apparatus (BD Biosciences, **Fig. S1B**). Sorted $CD8^+$ T cells were plated and incubated at 37°C for 48 hours in the presence of irradiated splenocytes (4,000 cGy) from syngeneic mice pulsed with the relevant peptide (4 μM for the ELISpot assay and 10^{-4} to 10^{-14}

M for the avidity assay). As a negative control, $CD8^+$ T cells from naive mice were incubated with peptide-pulsed splenocytes. Spots were revealed using the reagent set manufacturer protocol and were enumerated using an ImmunoSpot S5 UV Analyzer (Cellular Technology Ltd). IFN-γ production was expressed as the number of spot-forming cells per 10^6 $CD8^+$ T cells and the EC_{50} was calculated using a dose-response curve.

Cell isolation from lymphoid tissue and tetramer-based enrichment protocol. Spleen and lymph nodes (inguinal, axillary, brachial, cervical and mesenteric) were harvested from naive C57BL/6 mice or at day 0 for immunized mice, at signs of disease or day 21 for non-immunized EL4-injected mice and at signs of disease or day 210 for rechallenged mice. Single-cell suspensions were stained with Fc block and 10 nM of PE- or APC-labeled peptide MHC (pMHC) tetramers (NIH Tetramer Core Facility) for 30 minutes at 4ºC. After washing with ice-cold sorting buffer (PBS with 2% FBS), cells were resuspended in 200 µL of sorting buffer and 50 µL of anti-PE and/or anti-APC antibody conjugated magnetic microbeads (Miltenyi Biotech), then incubated for 20 minutes at 4ºC. Cells were then washed and tetramer+ cells were magnetically enriched as previously described *(47, 48)*. The resulting $tetramer^+$-enriched fractions were stained with APC Fire 750-conjugated anti-B220, F4/80, CD19, CD11b, CD11c (BioLegend), PerCP-conjugated anti-CD4 (BioLegend), BV421-conjugated anti-CD3 (BD Biosciences), BB515-conjugated anti-$CD8^+$ (BD Biosciences), BV510-conjugated anti-CD44 (BD Biosciences) antibodies and Zombie NIR Fixable Viability Kit (BioLegend). Anti-CD11b and CD11c were left out for the analysis of post-immunization repertoires because these markers may be expressed by some activated T cells. The entire stained sample was then analyzed on a FACS CantoII cytometer (BD Biosciences) and fluorescent counting beads (Thermo Fisher

Scientific) were used to normalize results. As negative control, we enriched the antigen-specific $CD8^+$ T-cell repertoires targeting 3 viral epitopes: gp-33 from the lymphocytic choriomeningitis virus protein gp-33 (KAVYNFATC; H-2-D^b), M45 from the murine cytomegalovirus protein M45 (HGIRNASFI; H-2-D^b) and B8R from the vaccinia virus protein B8R (TSYKFESV; H-2-K^b).

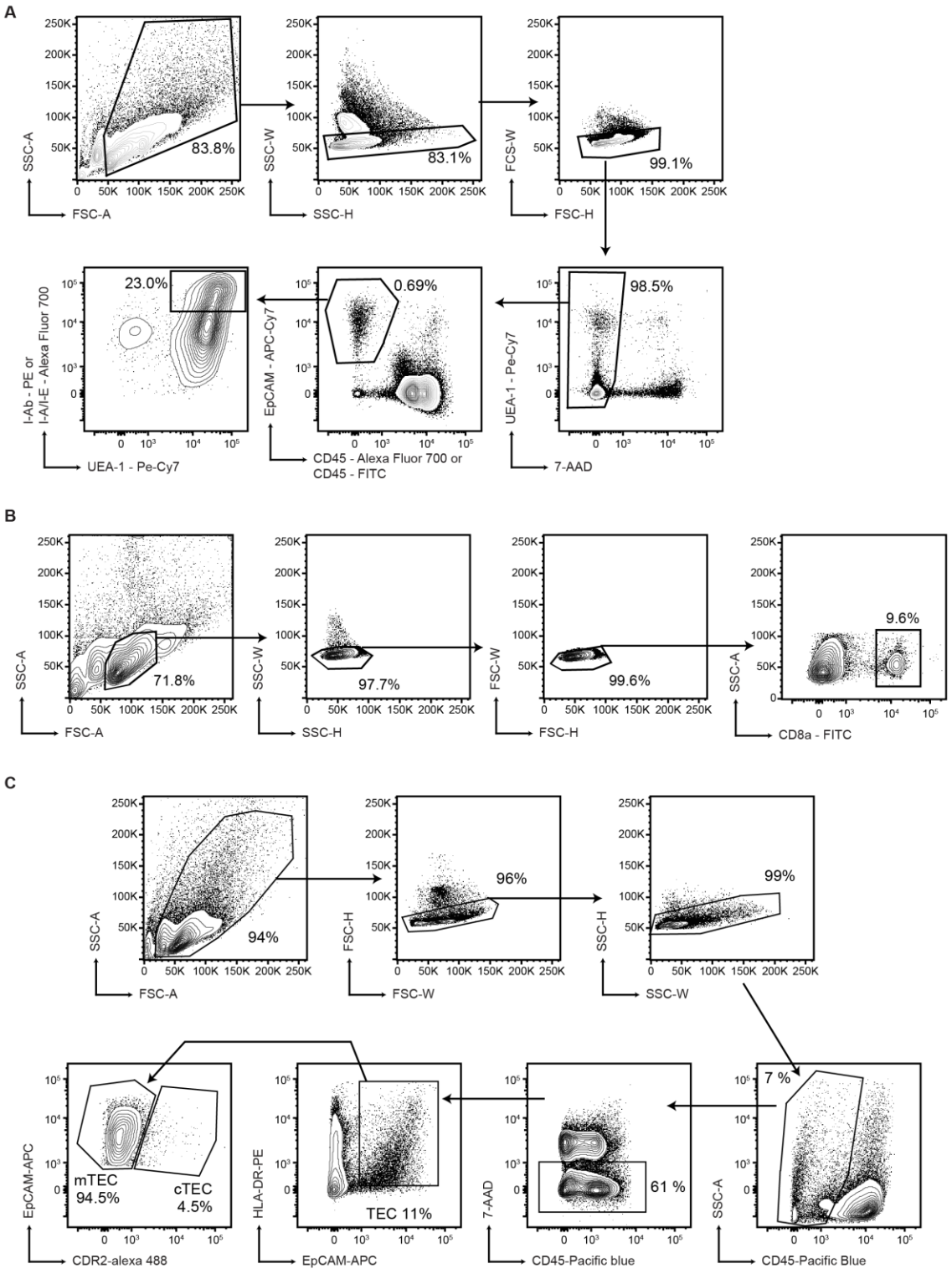


Fig. S1. Gating strategies for cells isolated by fluorescence-activated cell sorting. (**A**) Gating strategy for the isolation of murine mTEC^{hi}. mTEC^{hi} isolation was performed on single-cell

suspensions isolated from thymi of C57BL/6 or Balb/c mice. After doublets exclusion, $mTEC^{hi}$ cells were defined as $7\text{-}AAD^{-}$, $EpCAM^{+}$, $CD45^{-}$ (Alexa Fluor 700 for C57BL/6 or FITC for Balb/c mice), UEA-1^{+} and I-Ab^{+} (C57BL/6 mice) or I-A/I-E^{+} (Balb/c mice). (**B**) Gating strategy for the isolation of $CD8^{+}$ T cells for IFN-γ ELISpot assays. $CD8^{+}$ T cell isolation was performed on single-cell suspensions isolated from the spleen of naive or immunized C57BL/6 mice. After doublets exclusion, the CD8a marker was used to enrich for $CD8^{+}$ T cells. (**C**) Gating strategy for the isolation of human TECs and mTECs. Cell sorting was performed on single-cell suspensions isolated from thymi that were obtained from 3-month-old to 7-year-old individuals undergoing corrective cardiovascular surgery. After doublets exclusion, TECs were defined as $CD45^{-}$, $7\text{-}AAD^{-}$, $EpCAM^{+}$ and HLA-DR^{+}. mTECs were further defined as $CDR2^{-}$.

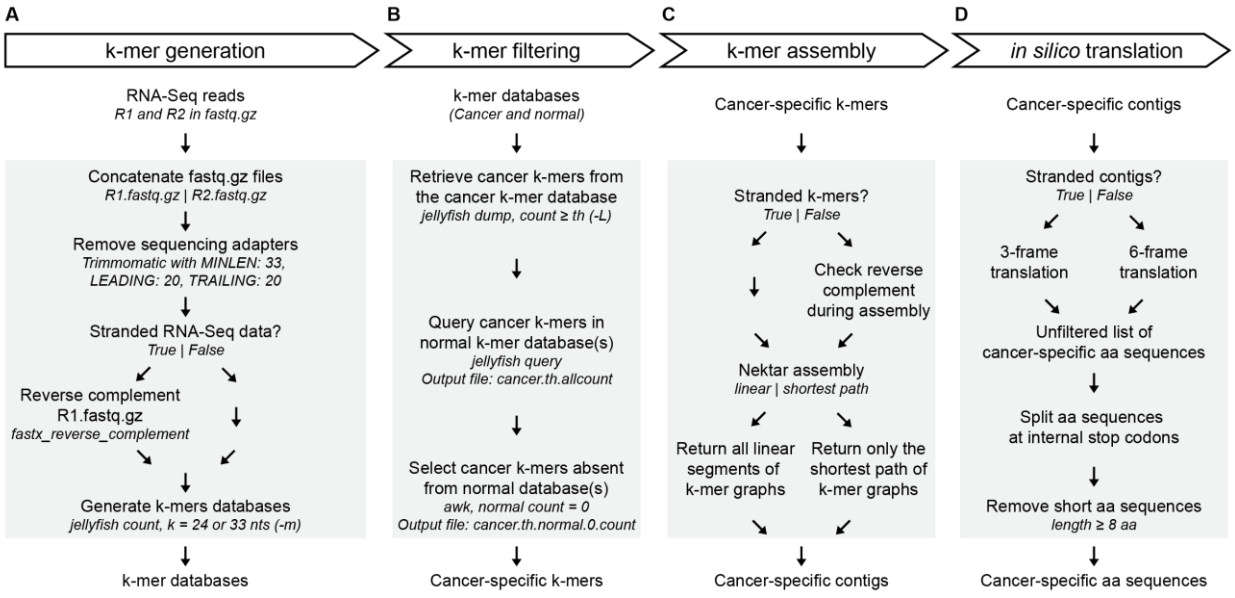


Fig. S2. Architecture of the codes used for our k-mer profiling workflow. (**A-D**) Details pertaining to the codes used to generate k-mers from RNA-Seq reads (**A**), filter k-mers (**B**), assemble k-mers into contigs (**C**) and translate contigs (**D**).

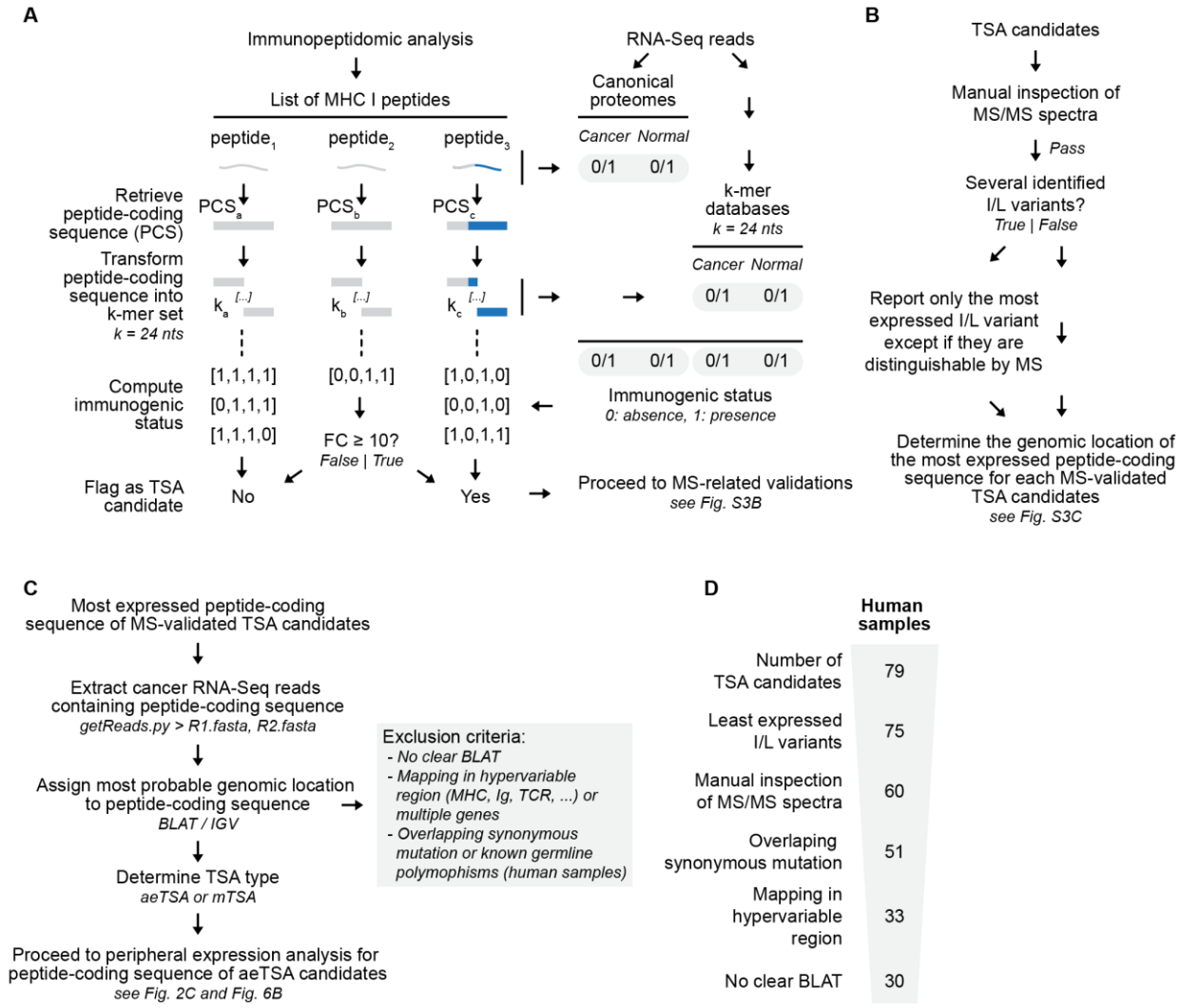
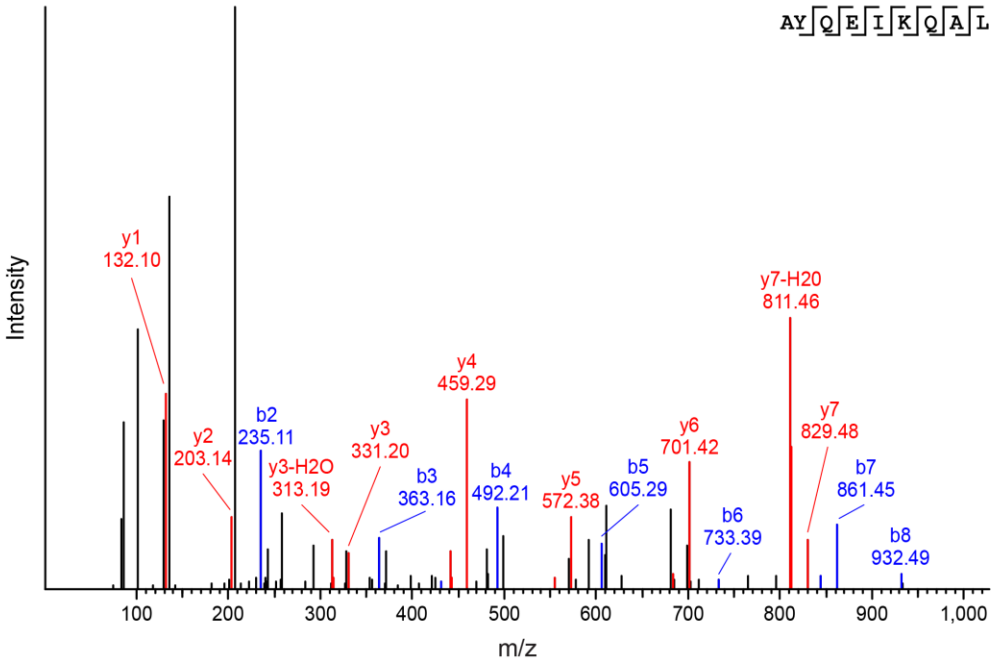


Fig. S3. TSA validation process. (**A**) Schematic detailing the computational strategy used for identification of TSA candidates. FC: Tumor / syngeneic mTEC^{hi} (murine samples) or TEC/mTEC (human samples). (**B**) Strategy used to perform the MS-related validations of MHC peptides flagged as TSA candidates. (**C**) Schematic summarizing the strategy used to assign a genomic location to MS-validated murine TSA candidates (CT26 and EL4) as well as MS-validated human TSA candidates for B-ALL specimens and lung cancers. (**D**) Flowchart indicating key steps involved in human TSA validation, where the number of remaining peptides after each validation step is indicated (see **Fig. S3, A-C** for details).

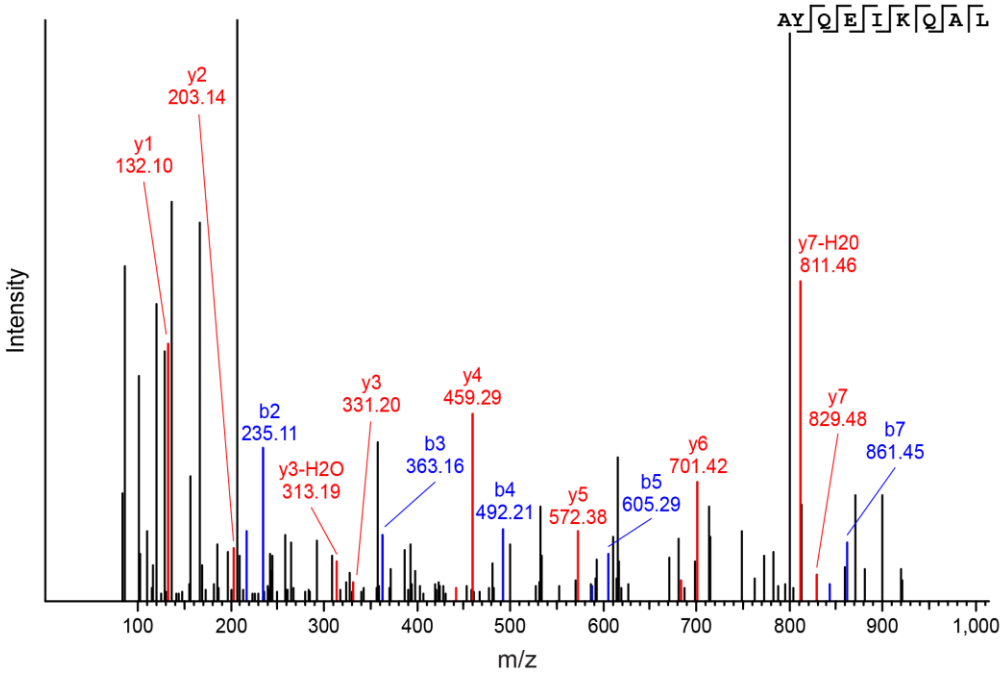
A

AYQEIKQAL - ERE aeTSA

Synthetic peptide



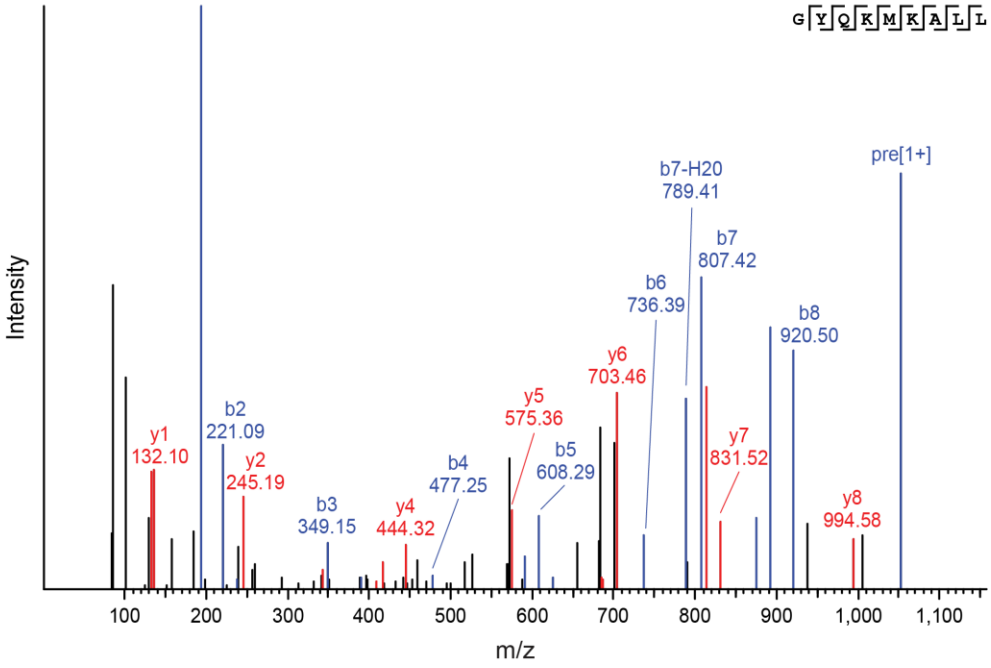
Endogenous peptide



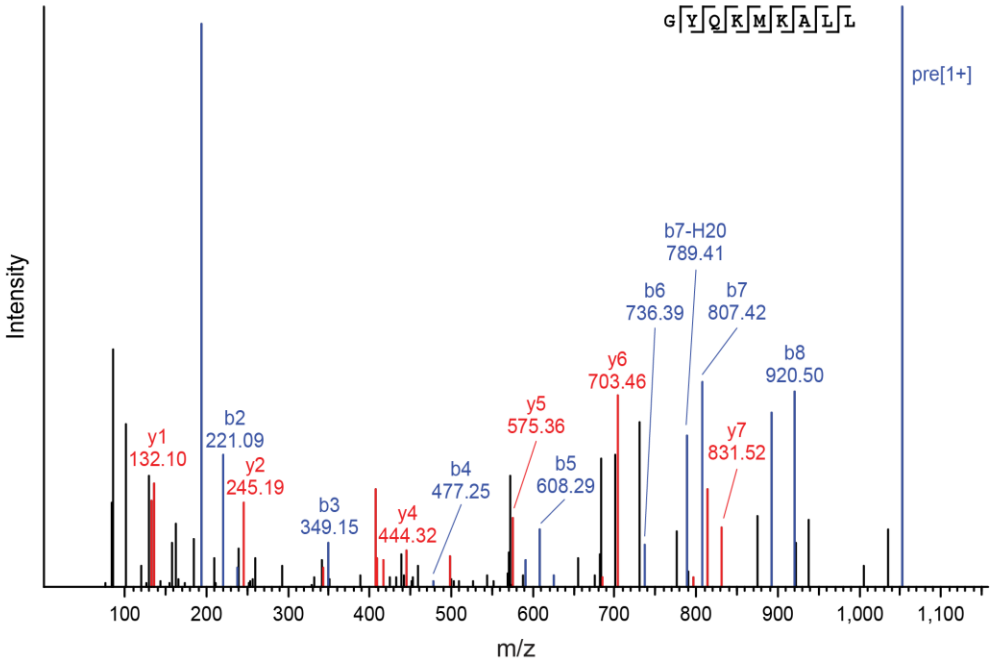
B

GYQKMKALL - ERE aeTSA

Synthetic peptide



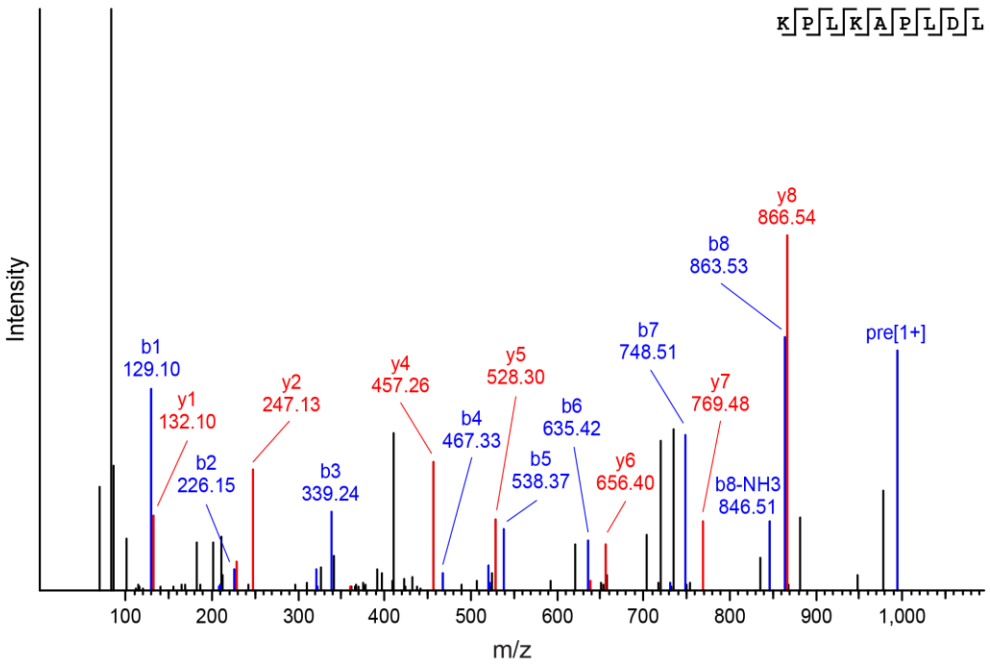
Endogenous peptide



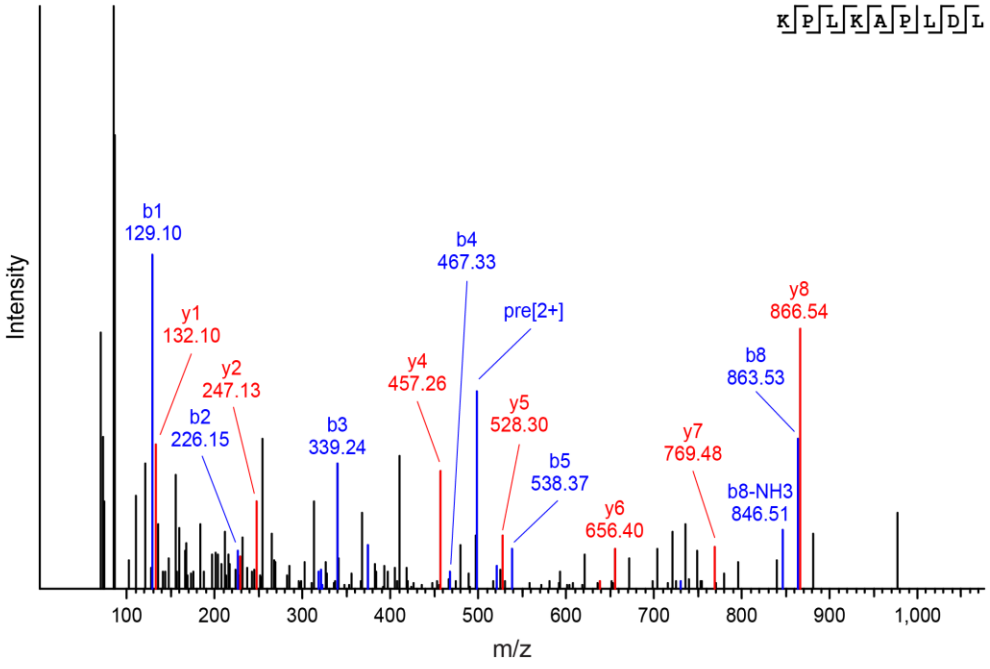
C

KPLKAPLDL - mTSA

Synthetic peptide



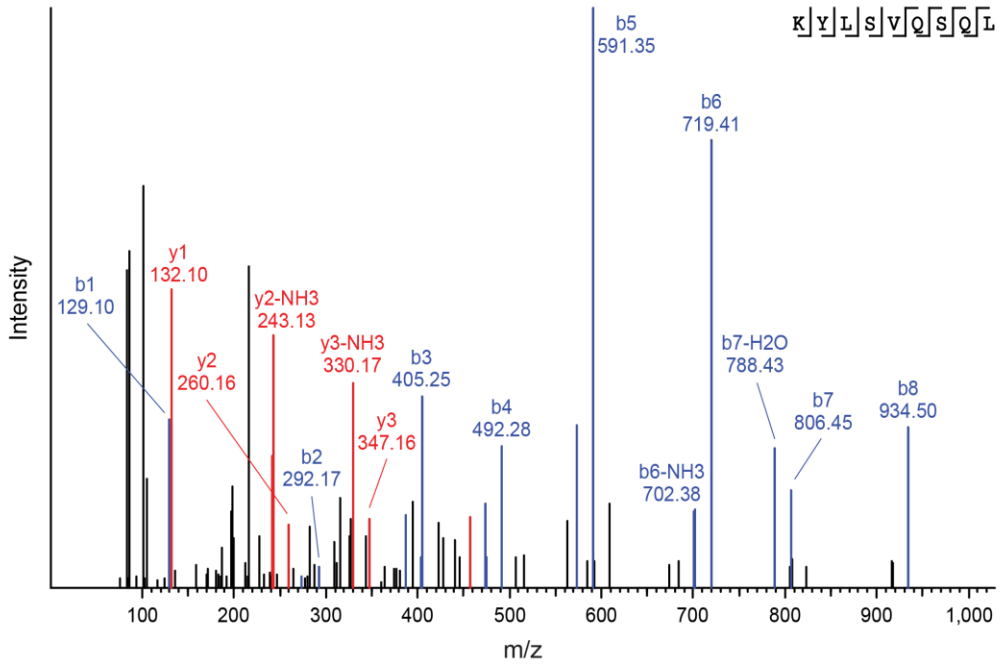
Endogenous peptide



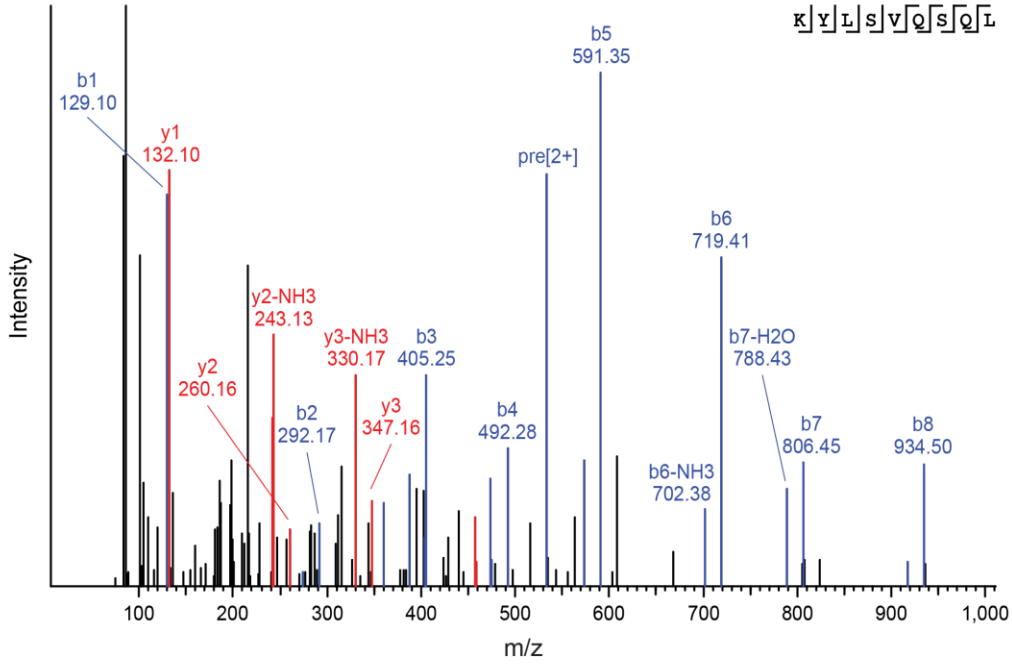
D

KYLSVQSQL - mTSA

Synthetic peptide



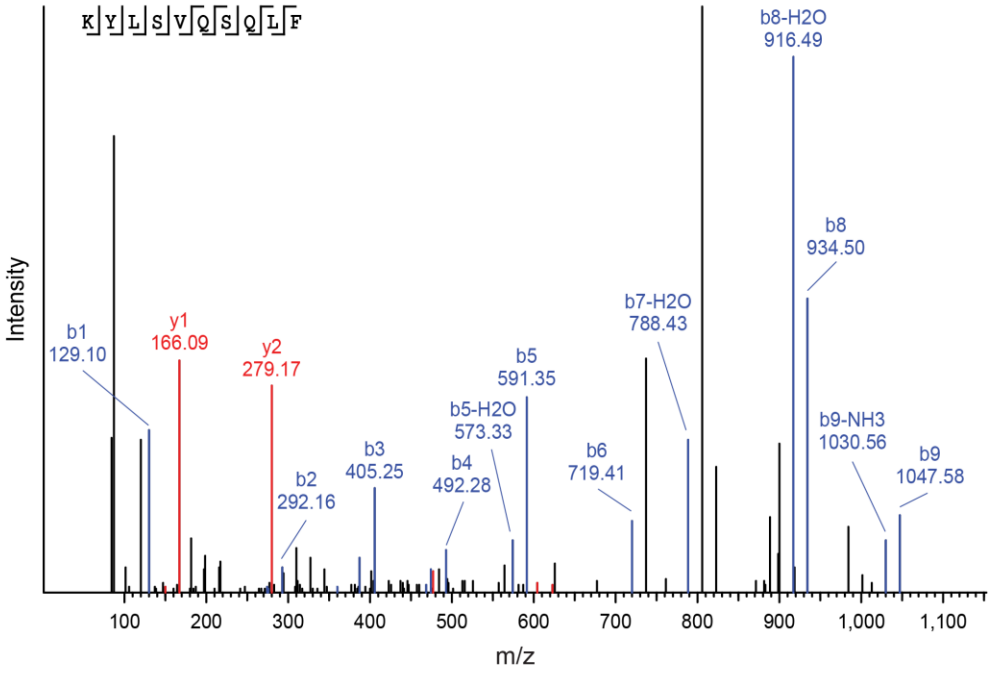
Endogenous peptide



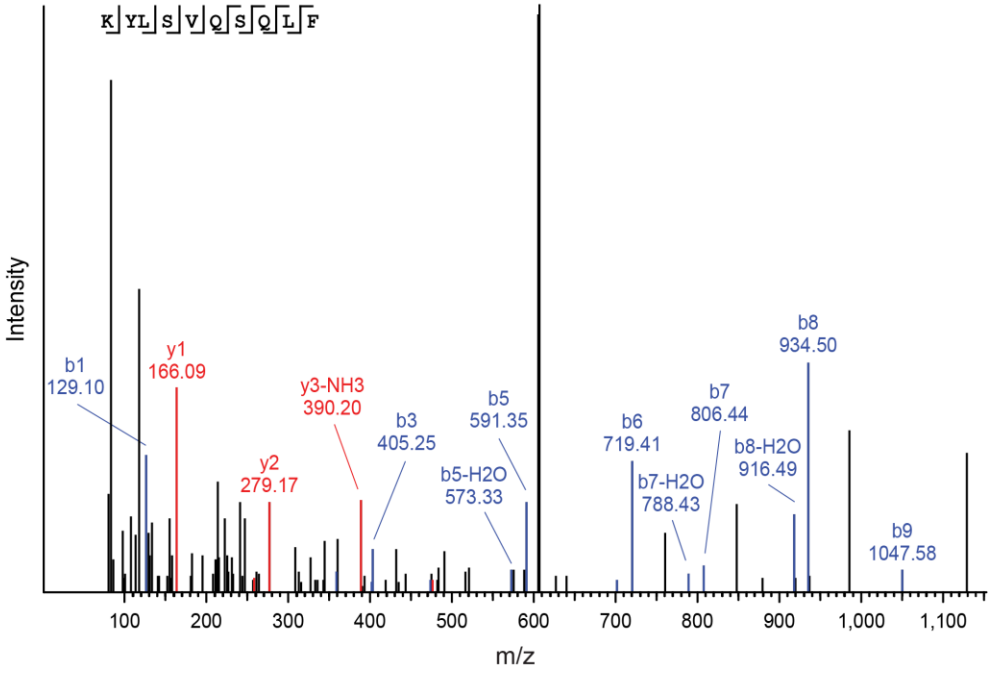
E

KYLSVQSQLF - mTSA

Synthetic peptide



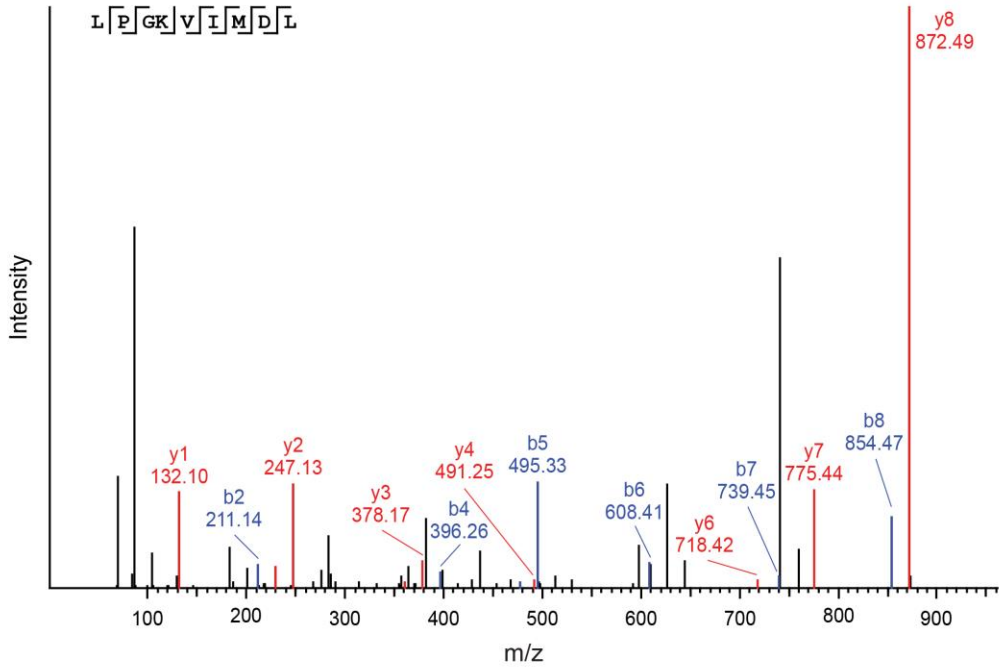
Endogenous peptide



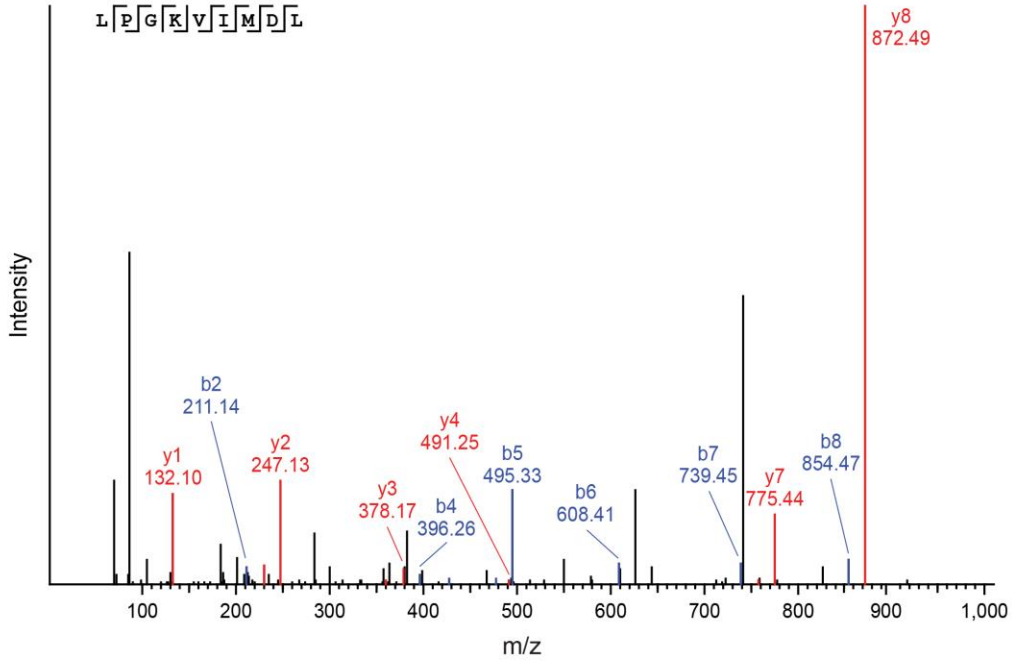
F

LPGKVIMDL - aeTSA

Synthetic peptide



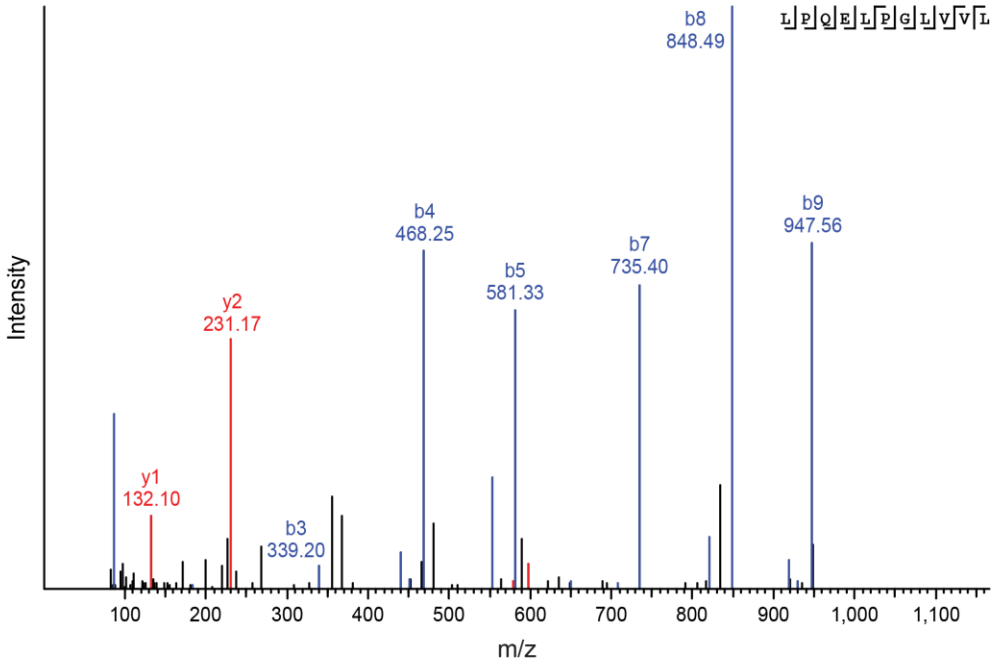
Endogenous peptide



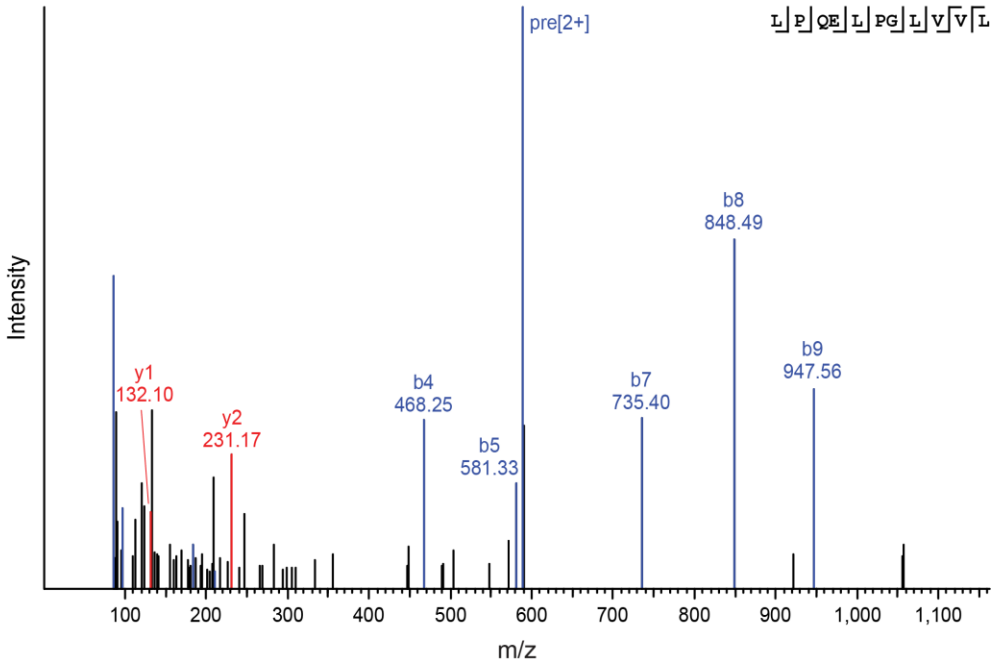
G

LPQELPGLVVL - ERE aeTSA

Synthetic peptide



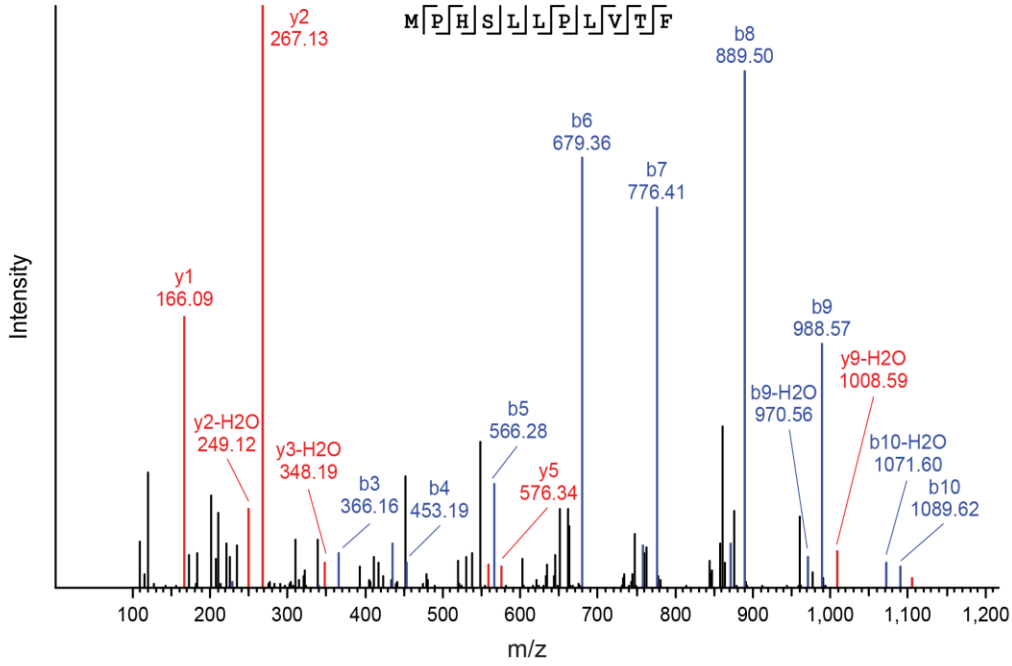
Endogenous peptide



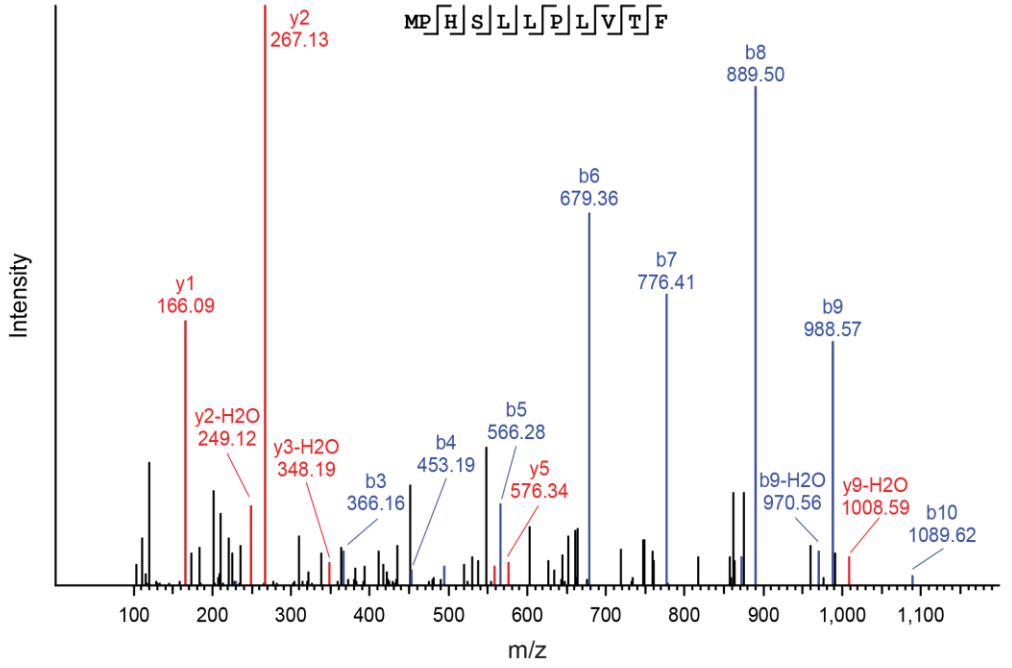
H

MPHSLLPLVTF - aeTSA

Synthetic peptide



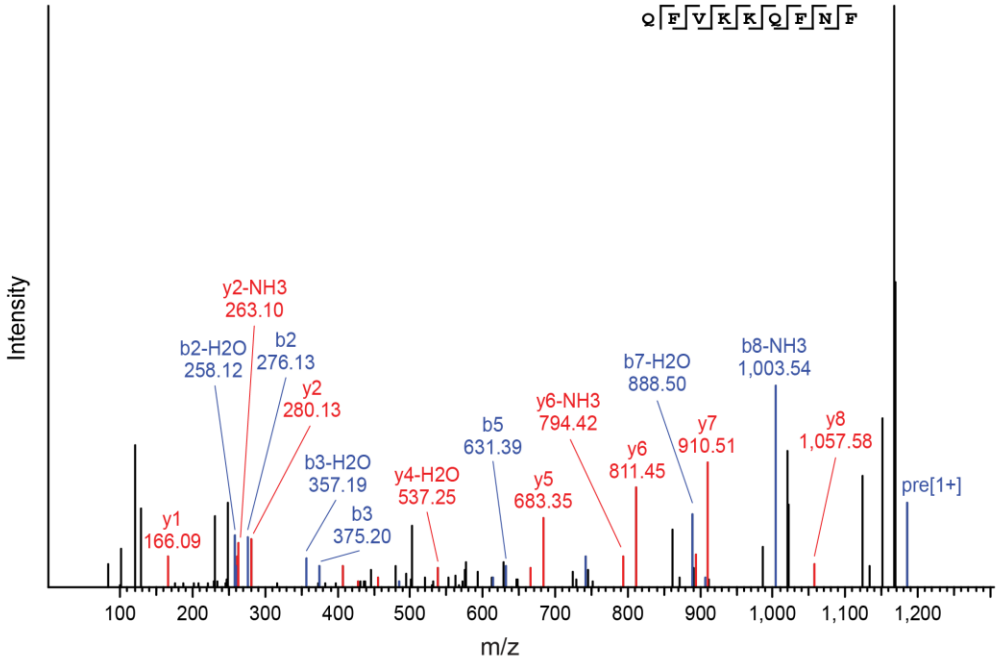
Endogenous peptide



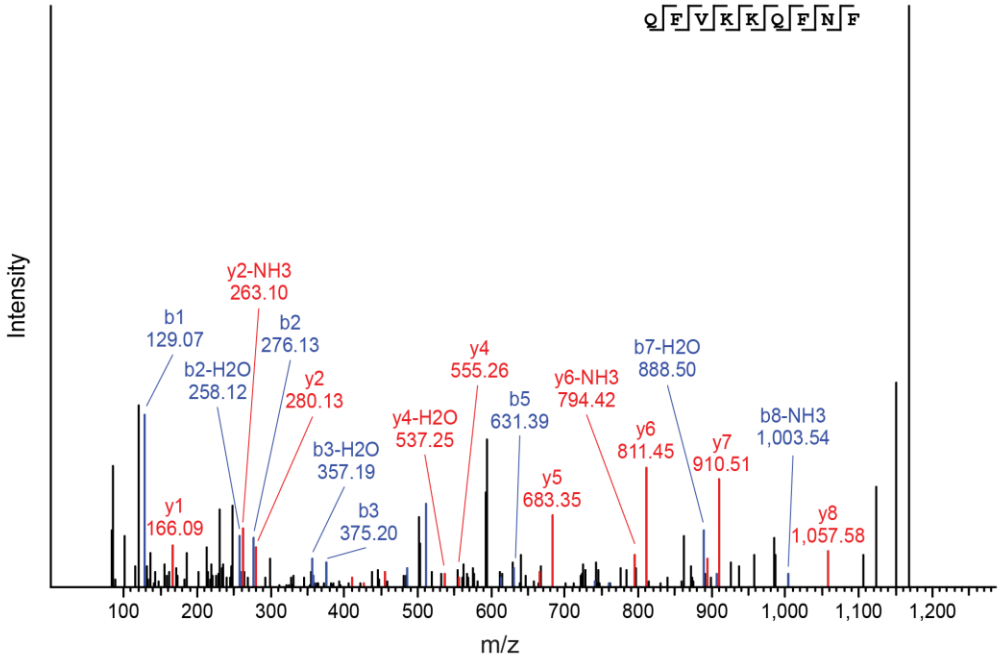
I

QFVKKQFNF - aeTSA

Synthetic peptide



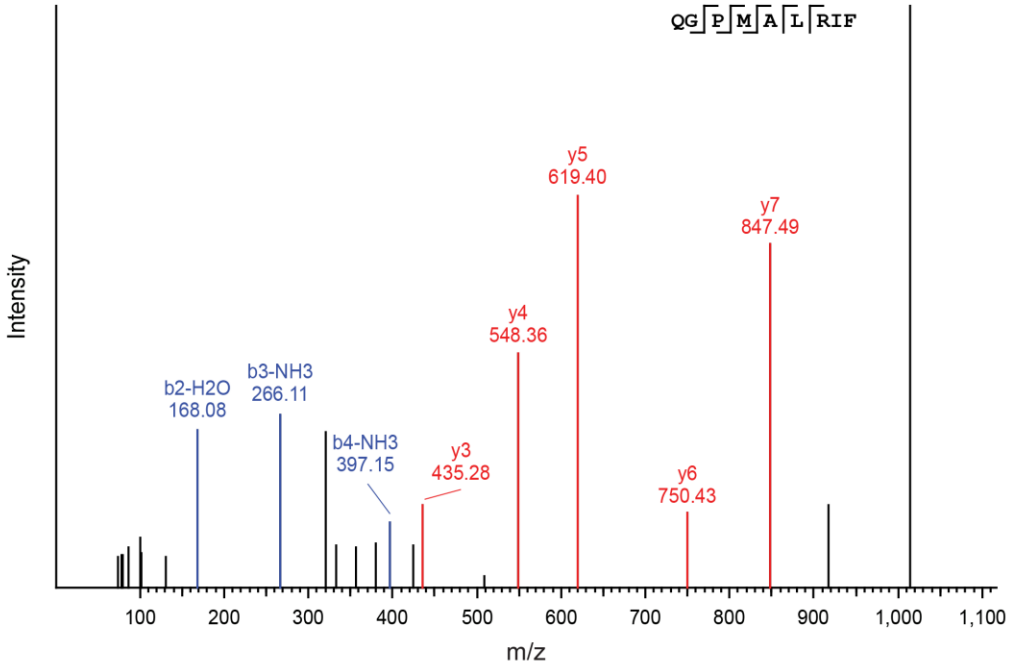
Endogenous peptide



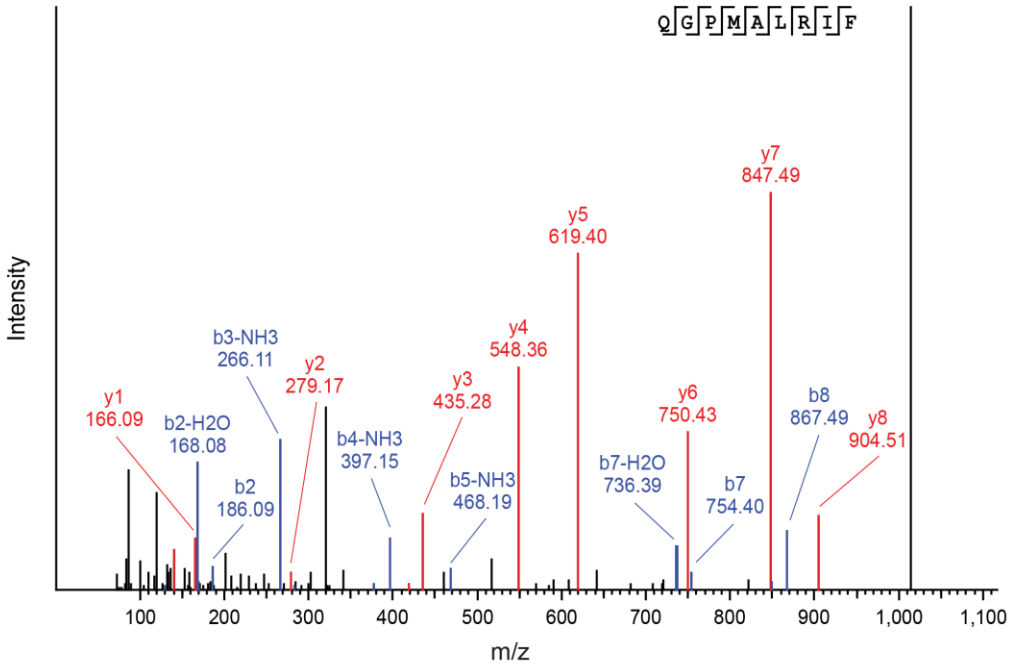
J

QGPMALRIF - mTSA

Synthetic peptide



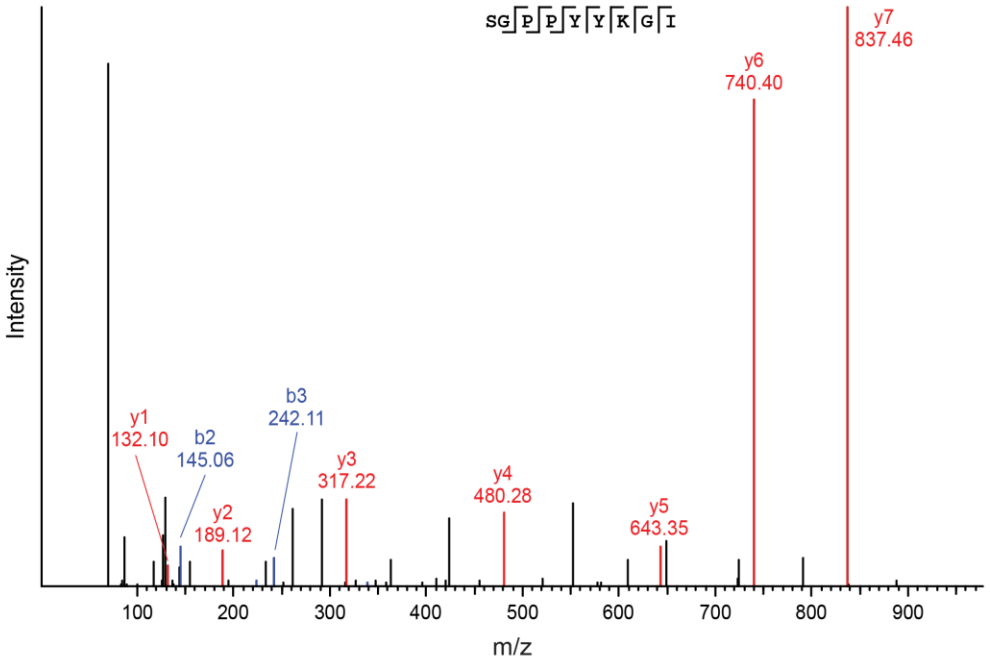
Endogenous peptide



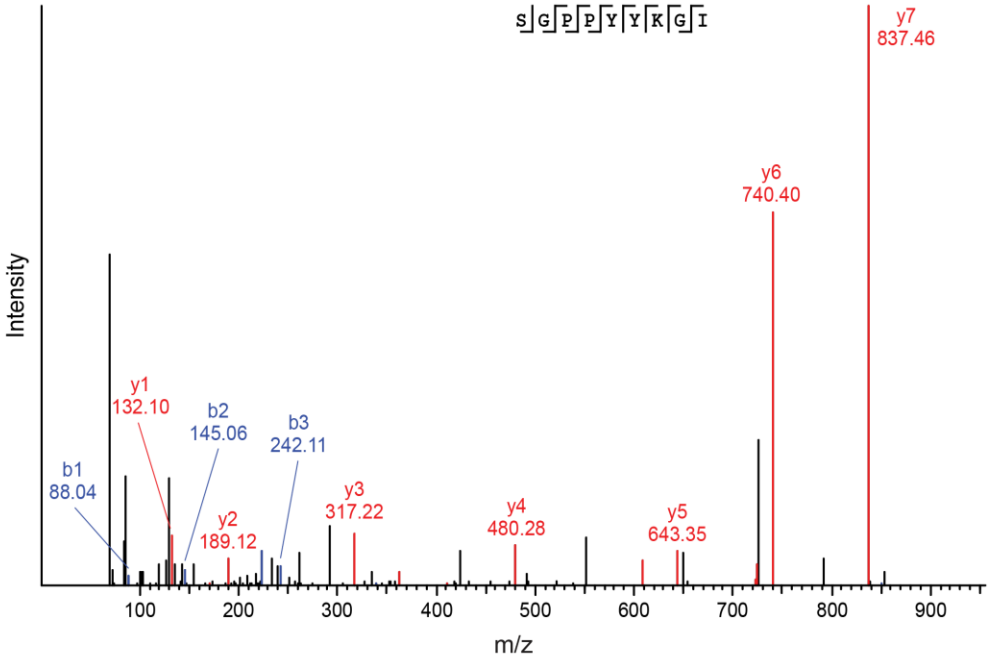
K

SGPPYYKGI - ERE aeTSA

Synthetic peptide



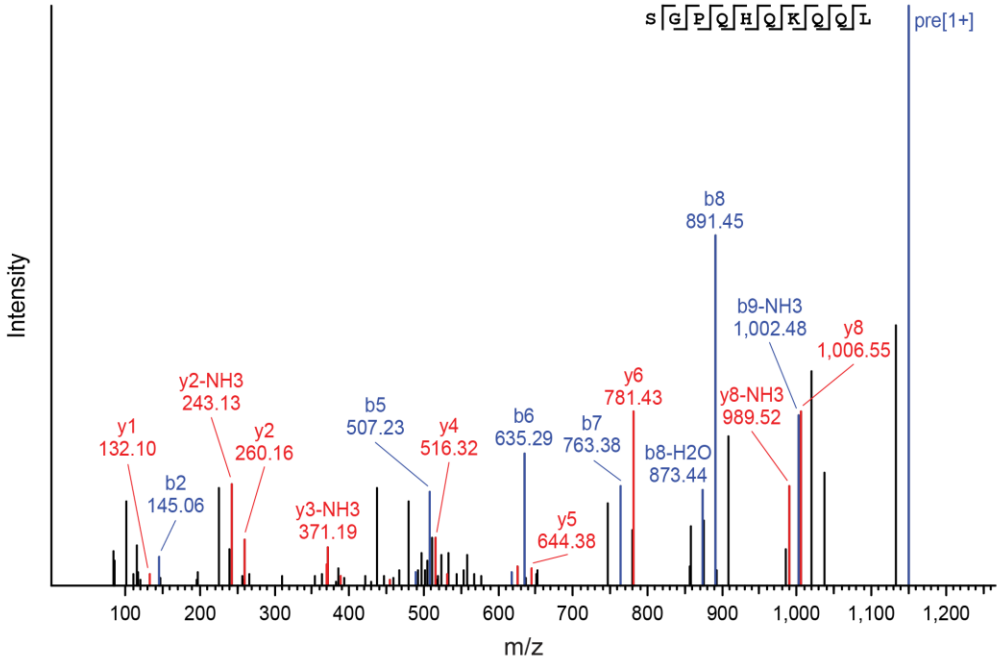
Endogenous peptide



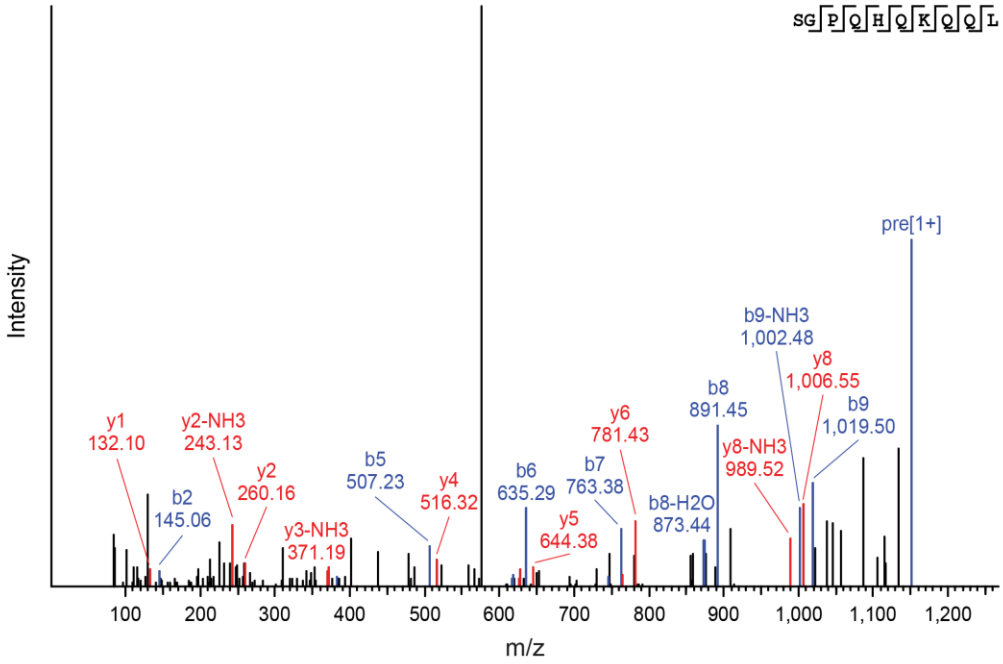
L

SGPQHQKQQL - aeTSA

Synthetic peptide



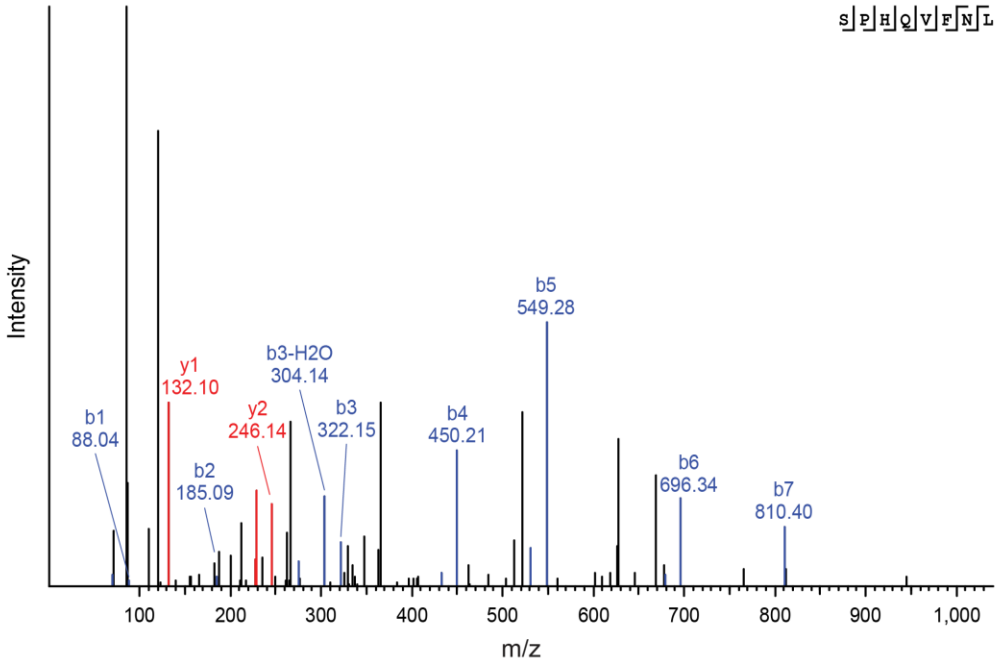
Endogenous peptide



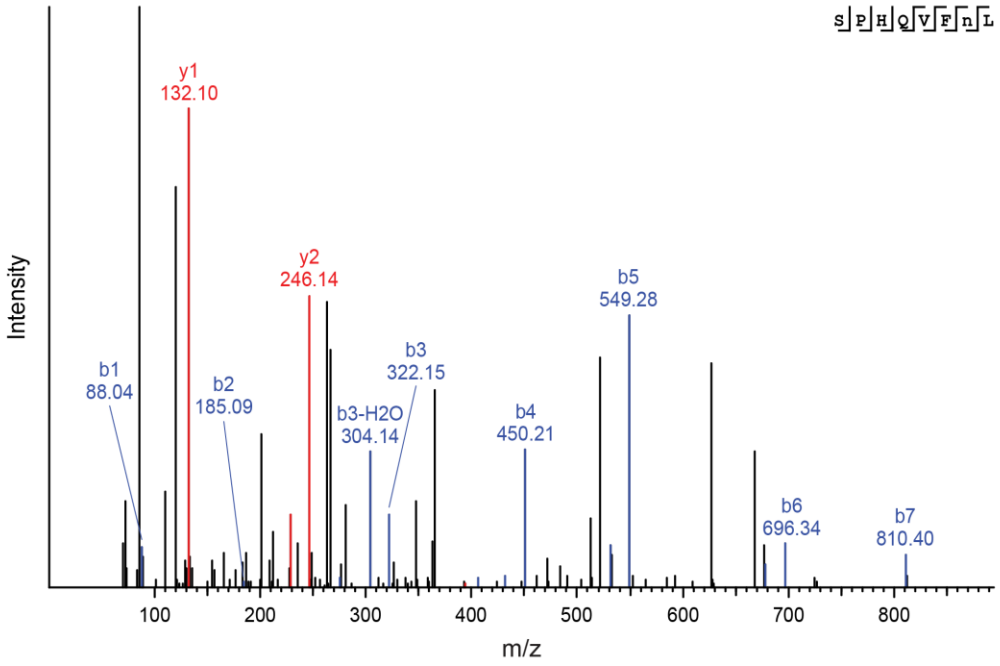
M

SPHQVFNL - ERE aeTSA

Synthetic peptide



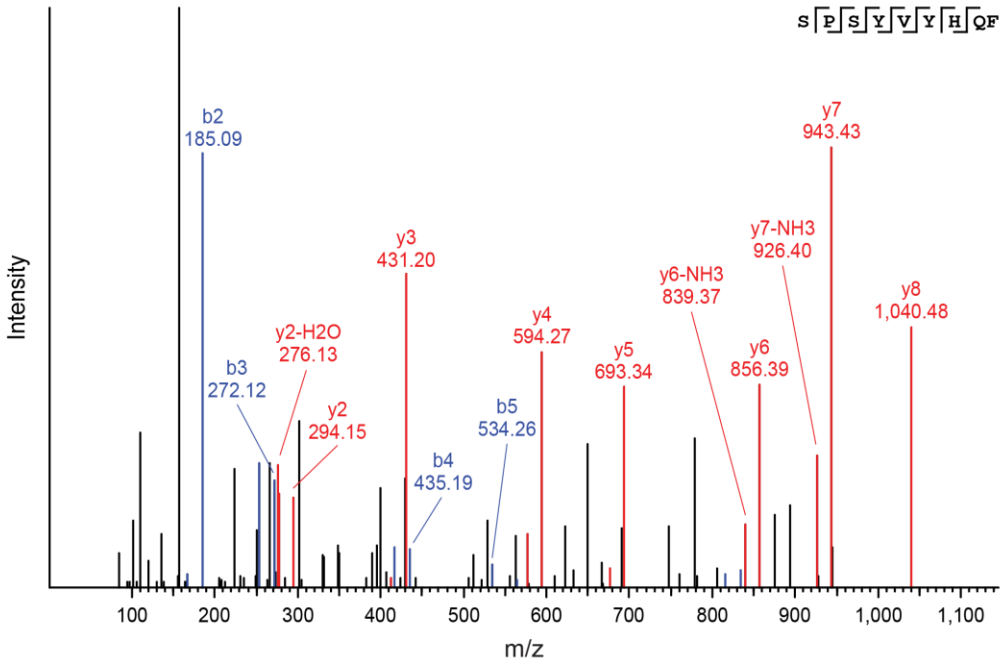
Endogenous peptide



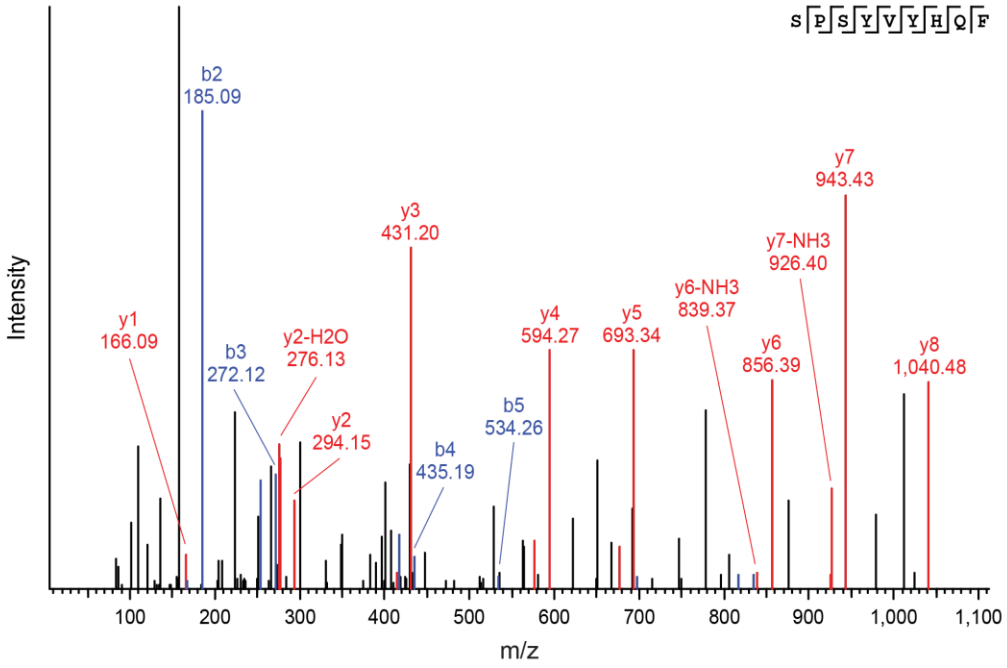
N

SPSYVYHQF - ERE aeTSA

Synthetic peptide



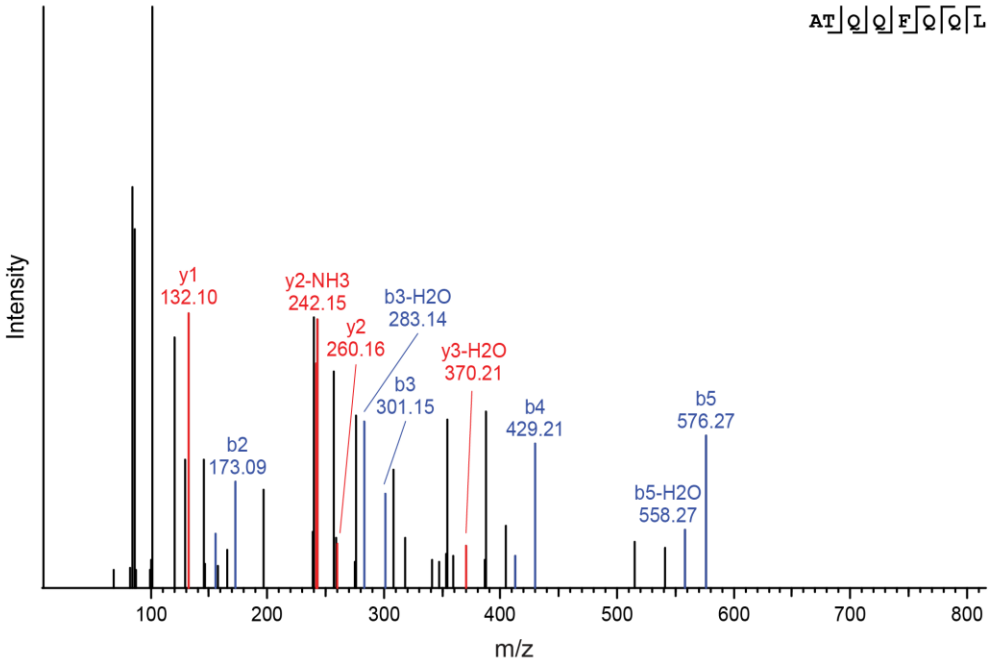
Endogenous peptide



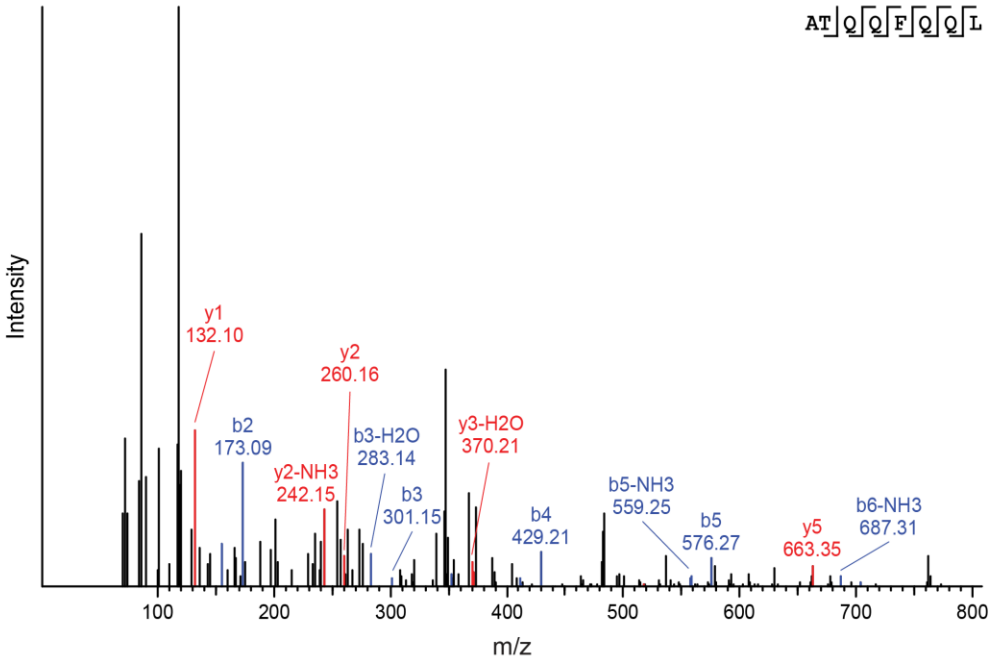
O

ATQQFQQL - ERE aeTSA

Synthetic peptide



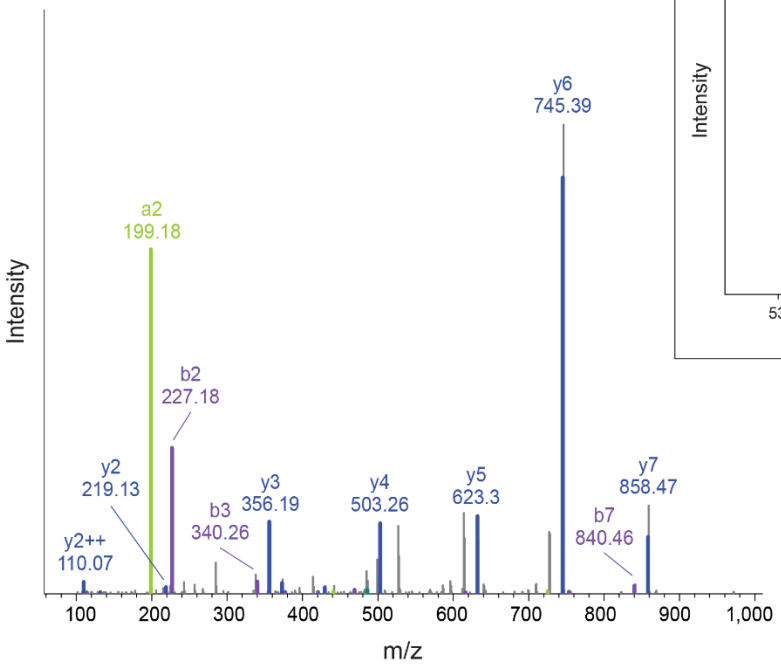
Endogenous peptide

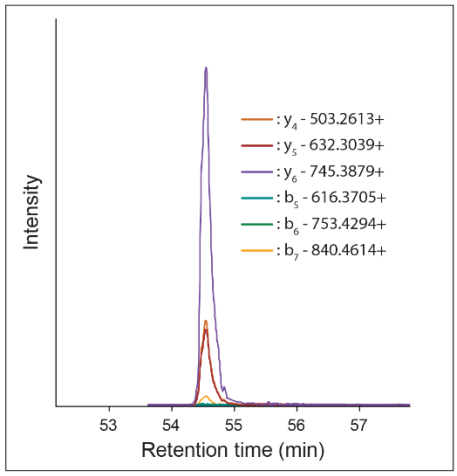


P

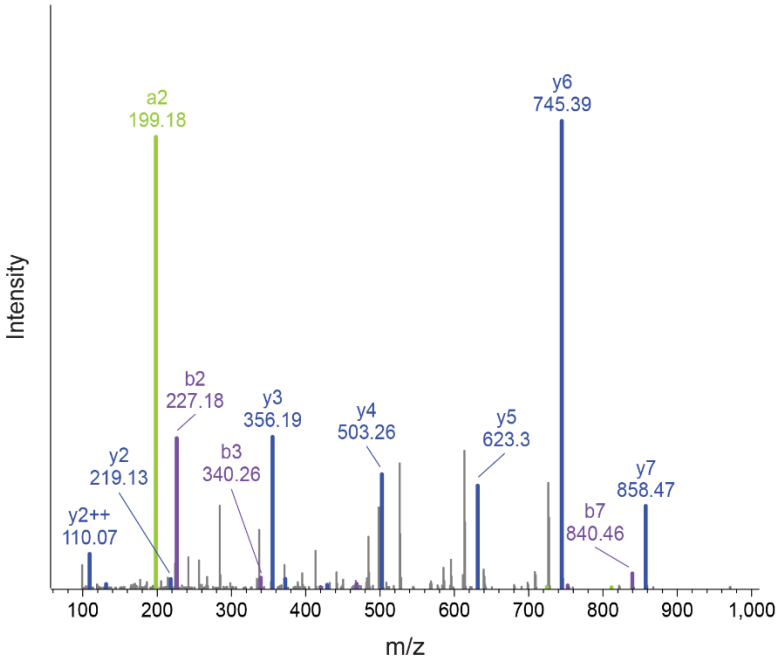
IILEFHSL - aeTSA

Synthetic peptide





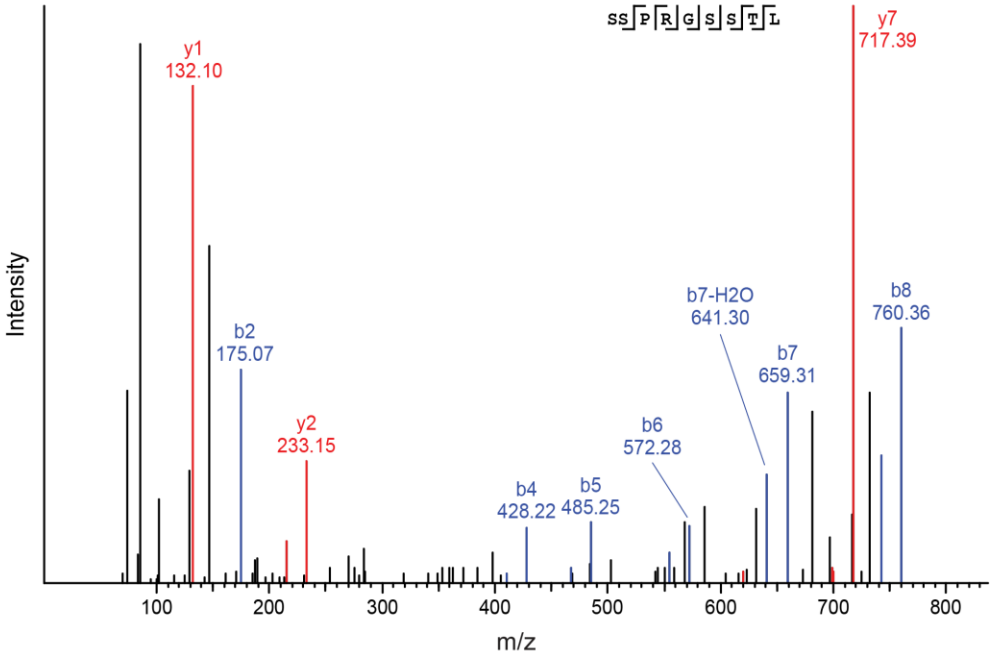
Endogenous peptide



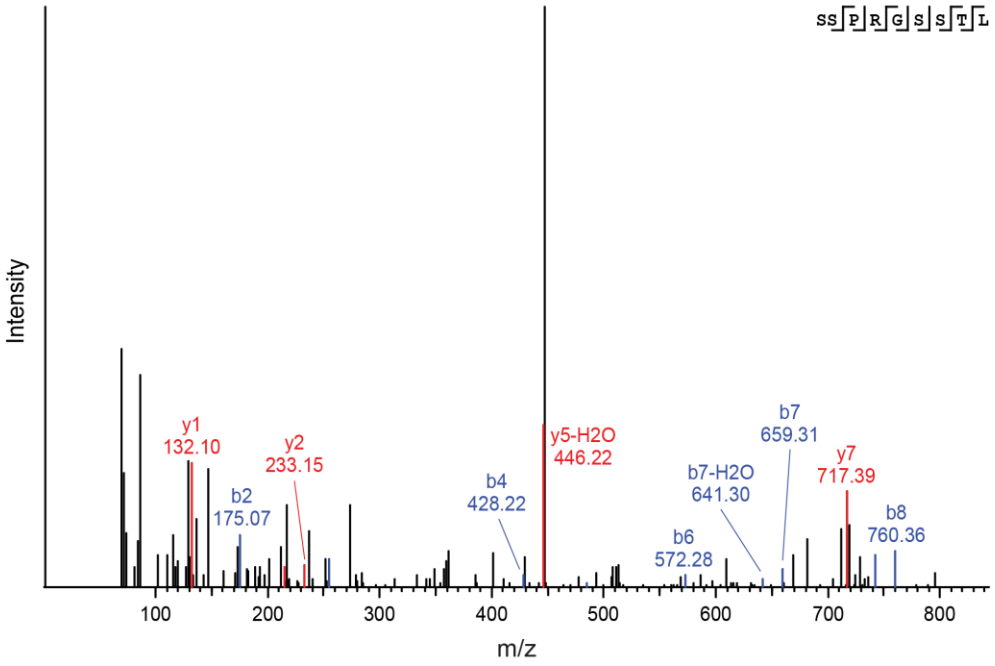
Q

SSPRGSSTL - ERE aeTSA

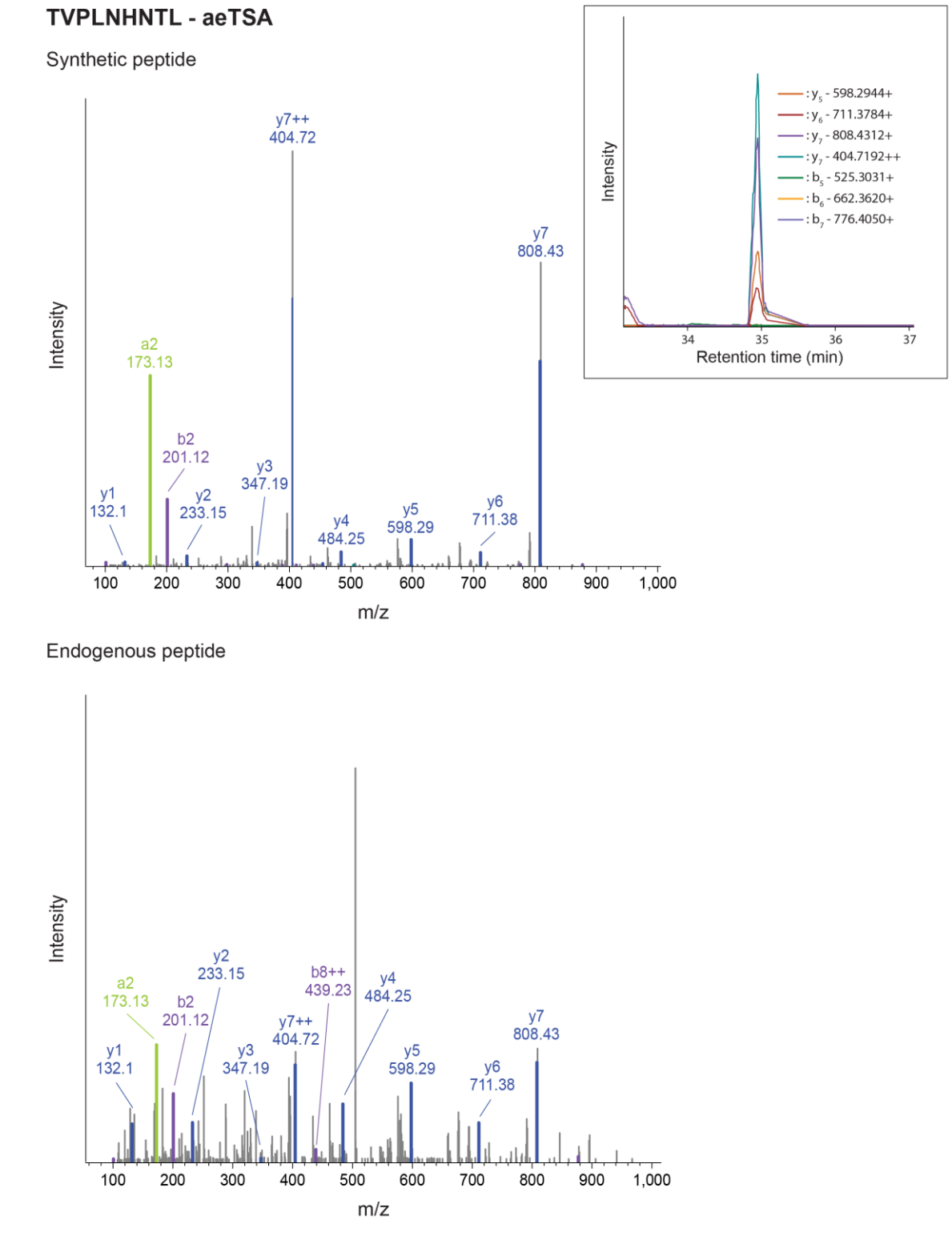
Synthetic peptide



Endogenous peptide



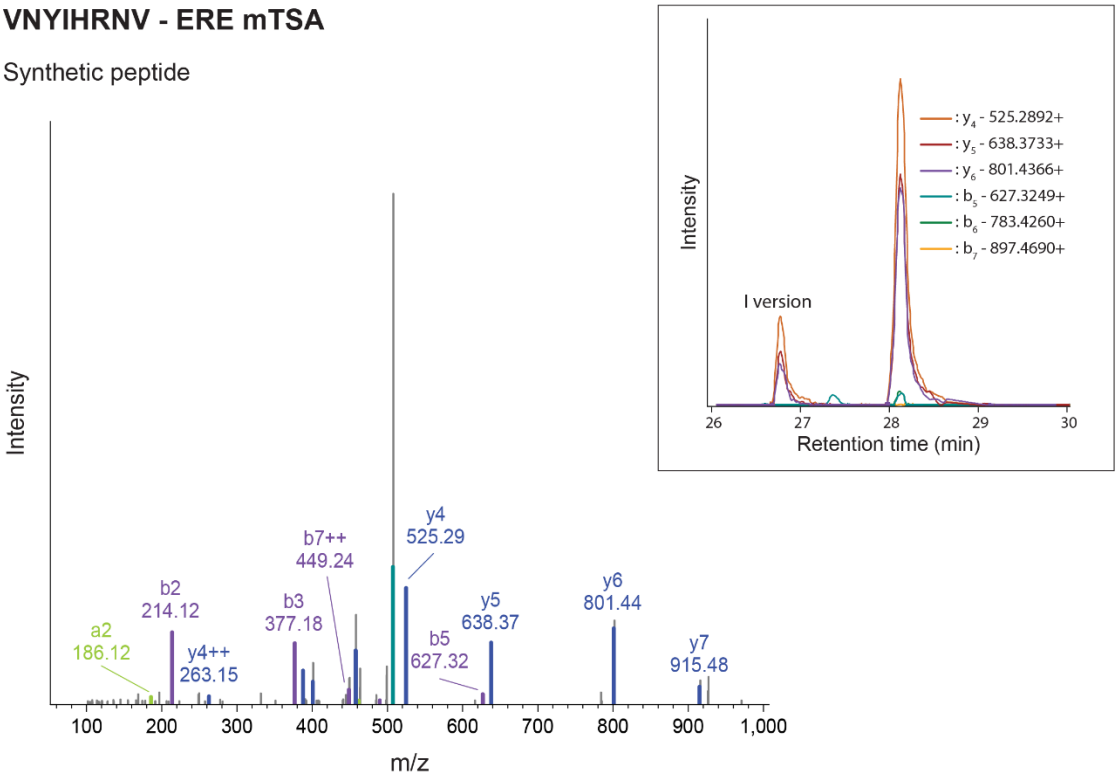
R


TVPLNHNTL - aeTSA
Synthetic peptide
Intensity
y7++
404.72
y7
808.43
a2
173.13
b2
201.12
y3
347.19
y1
132.1
y2
233.15
y4
484.25
y5
598.29
y6
711.38
100
200
300
400
500
600
700
800
900
1,000
m/z
Intensity
: y5 - 598.2944+
: y6 - 711.3784+
: y7 - 808.4312+
: y7 - 404.7192++
: b5 - 525.3031+
: b6 - 662.3620+
: b7 - 776.4050+
34
35
36
37
Retention time (min)
Endogenous peptide
Intensity
y2
233.15
b8++
439.23
y4
484.25
a2
173.13
b2
201.12
y7++
404.72
y7
808.43
y1
132.1
y3
347.19
y5
598.29
y6
711.38
100
200
300
400
500
600
700
800
900
1,000
m/z

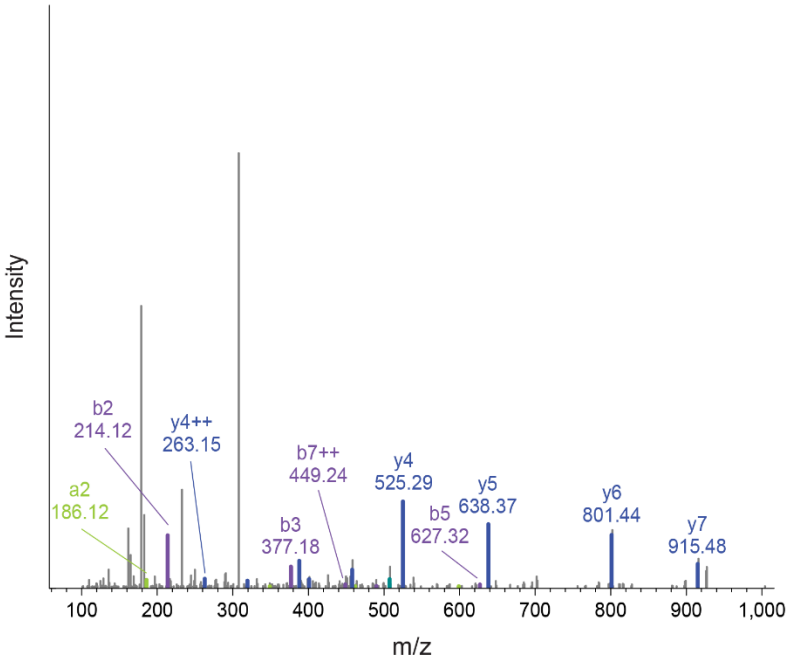
S

VNYIHRNV - ERE mTSA

Synthetic peptide



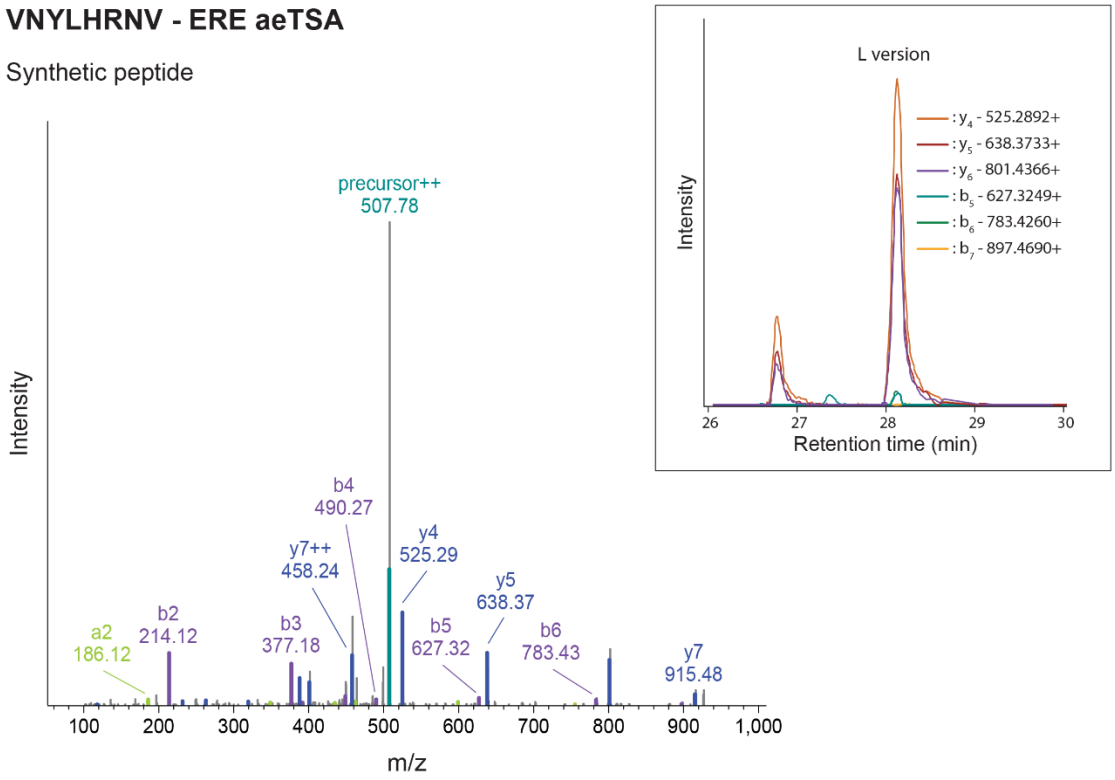
Endogenous peptide



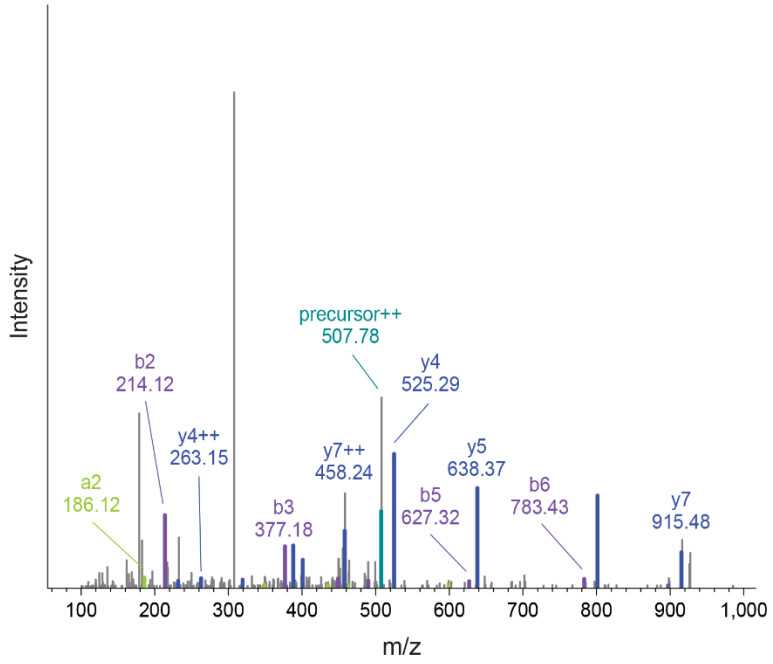
T

VNYLHRNV - ERE aeTSA

Synthetic peptide



Endogenous peptide



U

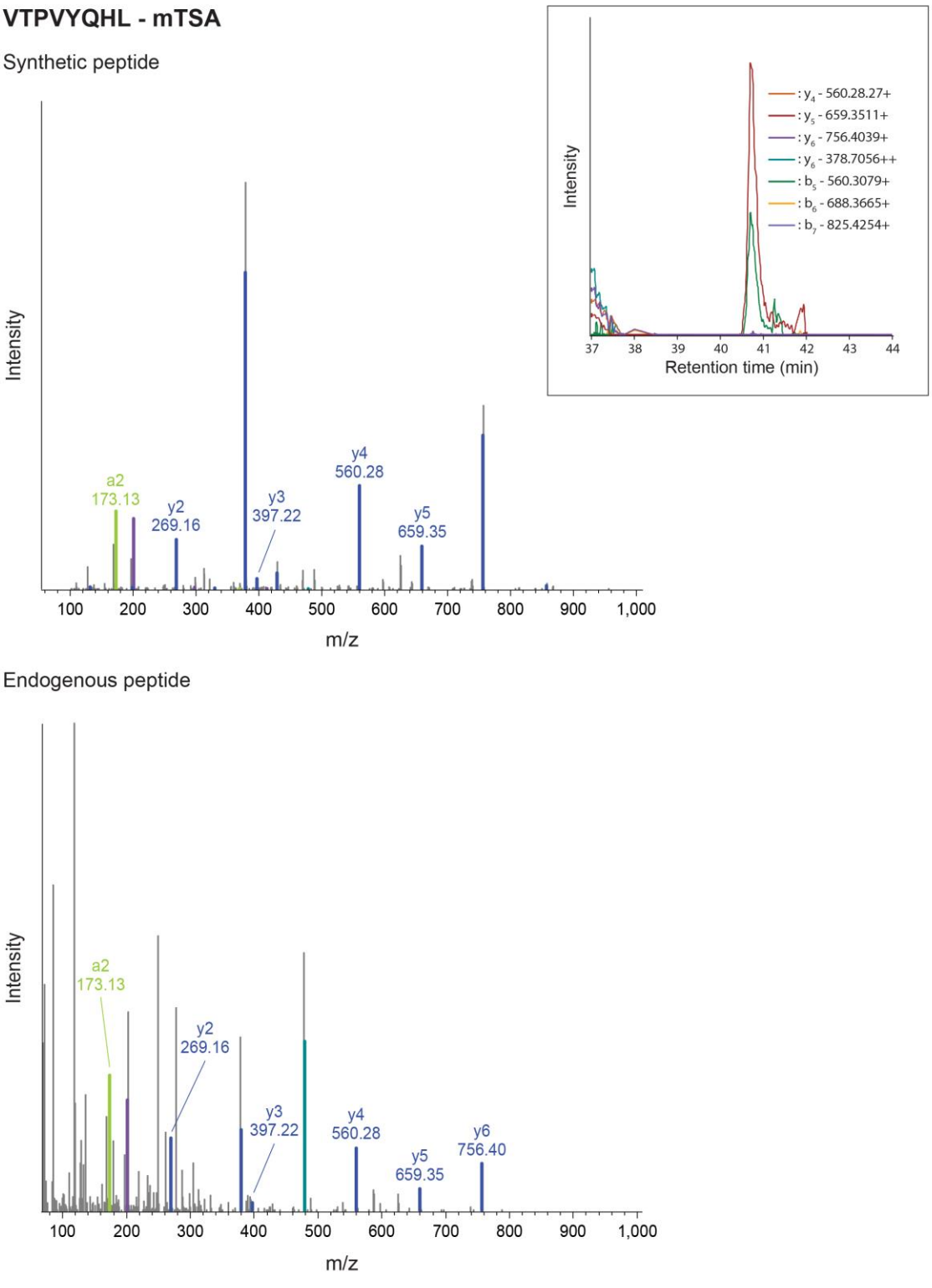

VTPVYQHL - mTSA
Synthetic peptide
Intensity
a2
173.13
y2
269.16
y3
397.22
y4
560.28
y5
659.35
100
200
300
400
500
600
700
800
900
1,000
m/z
: y4 - 560.28.27+
: y5 - 659.3511+
: y6 - 756.4039+
: y6 - 378.7056++
: b5 - 560.3079+
: b6 - 688.3665+
: b7 - 825.4254+
Intensity
37
38
39
40
41
42
43
44
Retention time (min)
Endogenous peptide
Intensity
a2
173.13
y2
269.16
y3
397.22
y4
560.28
y5
659.35
y6
756.40
100
200
300
400
500
600
700
800
900
1,000
m/z

Fig. S4. MS validation of CT26 and EL4 TSA candidates using synthetic analogs. (A-N) Synthetic and endogenous MS/MS spectra for CT26 TSA candidates. (**O-U**) Synthetic and endogenous MS/MS spectra for EL4 TSA candidates. Spectra presented in panels **P** and **R-U** come from additional EL4 replicates analyzed by PRM MS. See section **MS validation of TSA candidates** of the **Supplementary Materials** for details.

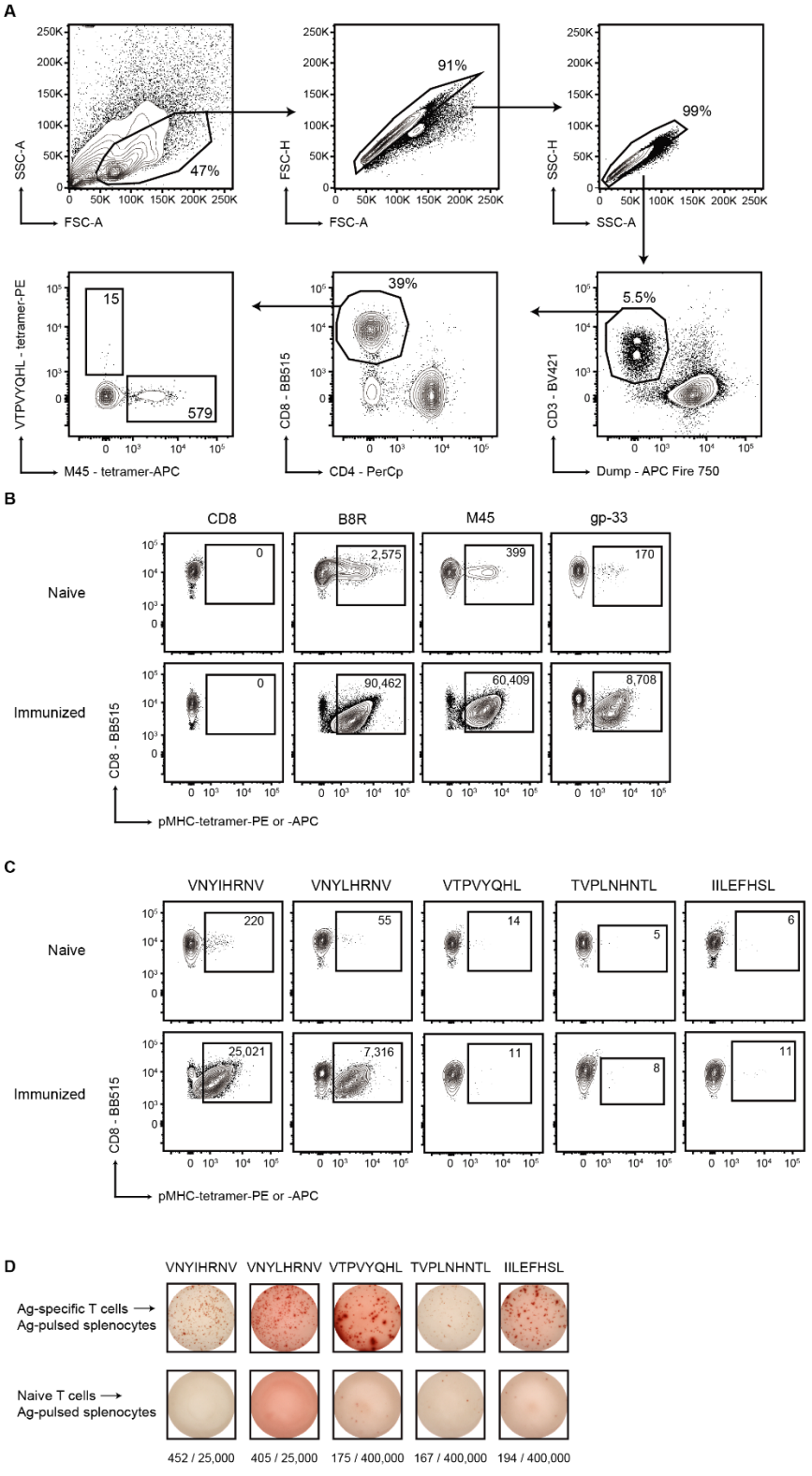


Fig. S5. Detection of antigen-specific CD8$^+$ T cells in naïve and immunized mice. (**A**) Gating strategy for the detection of tetramer$^+$ CD8$^+$ T cells ex vivo. Tetramer enrichment were

performed on single-cell suspensions isolated from the spleen and lymph nodes of each mice. After doublets exclusion, Dump$^-$ CD3$^+$ cells were analyzed for CD8 and CD4 expression and tetramer$^+$ cells were analyzed in the CD8$^+$ compartment. A representative staining obtained following VTPVYQHL/H-2-K^b-PE and M45/H-2-D^b-APC tetramers enrichment in a naive mouse is shown. Absolute numbers of tetramer$^+$ CD8$^+$ T cells detected for each specificity are indicated. The Dump channel corresponds to pooled events positive for 7-AAD, CD45R and CD19, F4/80, CD11b and CD11c. (**B** and **C**) Representative analysis of tetramer$^+$ CD8$^+$ T cells in naive (upper row) and immunized (lower row) mice. CD8$^+$ cells before magnetic enrichment and after ex vivo enrichment for tetramer$^+$ viral specificities (**B**) as well as for TSA specificities (**C**) are shown. Percentages and numbers of tetramer$^+$ or tetramer$^-$ cells are indicated. (**D**) One representative experiment of the frequency of IFN-γ secreting CD8$^+$ T cells in immunized and naive mice. The number of spot forming cells relative to the number of plated CD8$^+$ T cells in each condition are indicated below each well.

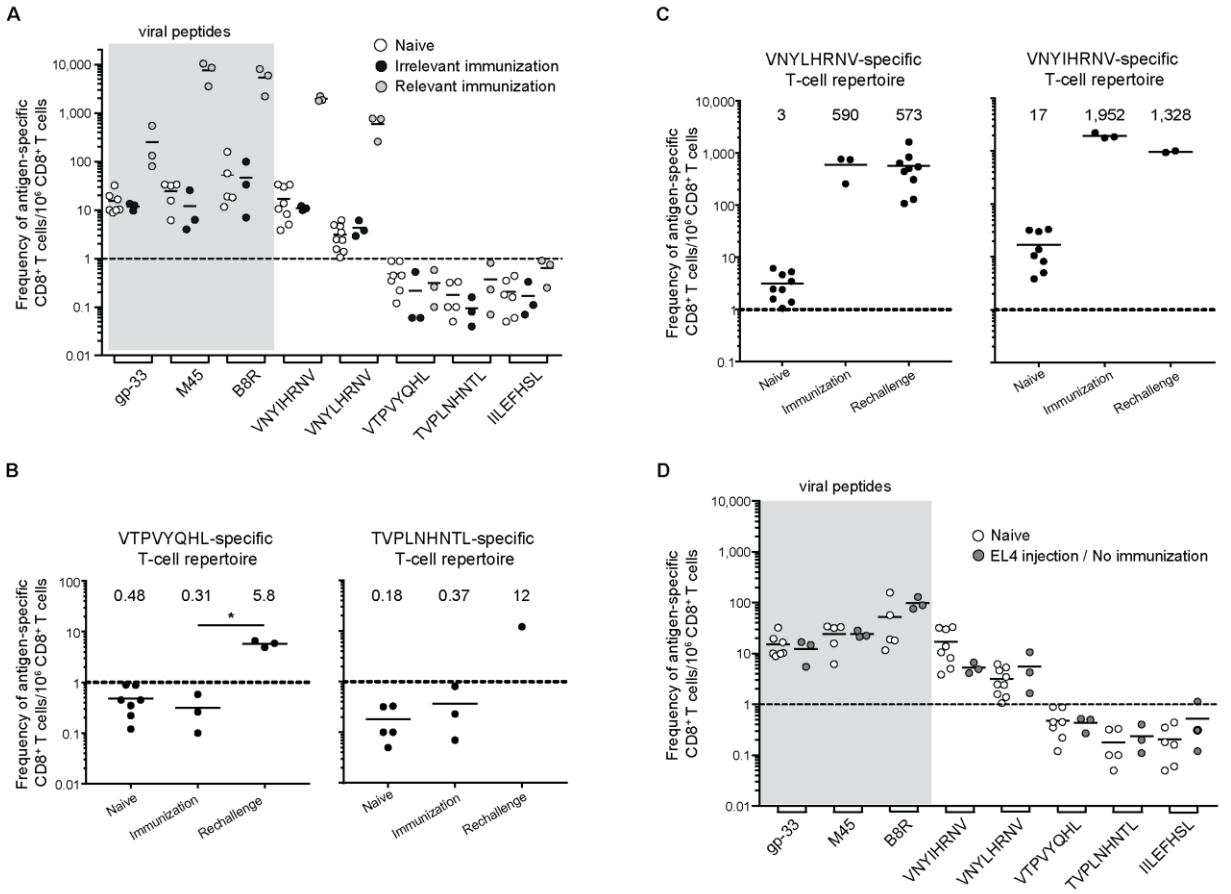


Fig. S6. Frequencies of antigen-specific T cells. (**A**) Frequencies of antigen-specific T cells in naïve mice and mice immunized with relevant or irrelevant peptides. (**B** and **C**) Frequencies of antigen-specific $CD8^+$ T cells in mice immunized against VTPVYQHL or TVPLNHNTL (**B**) or against VNYLHRNV or VNYIHRNV (**C**) that were rechallenged with EL4 cells at day 150. For comparison purposes, frequencies of antigen-specific T cells in naive and immunized mice reported in panel (**A**) are reproduced. (**D**) Frequencies of antigen-specific T cells in non-immunized mice injected with EL4 cells. All calculated frequencies of $tetramer^+$ $CD8^+$ T cells are expressed as the number of antigen-specific $CD8^+$ T cells per 10^6 $CD8^+$ T cell. Each symbol represents one mouse (n = 1 to 9 mice). Dotted line represents a minimal detection level of one $tetramer^+$ T cell per 10^6 $CD8^+$ T cells. Viral peptides used as controls are highlighted in gray. p-values were calculated using one-sided Wilcoxon rank sum test (* $p \leq 0.05$).

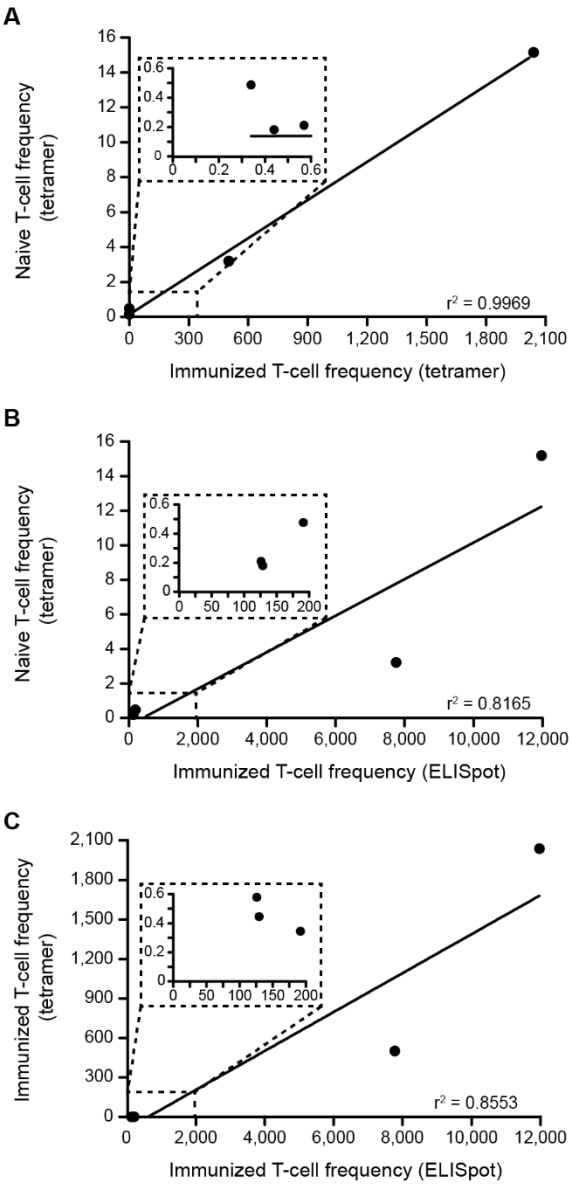


Fig. S7. Correlation between antigen-specific T cell frequencies in naïve and immunized mice. (**A** and **B**) Correlation between the frequencies of antigen-specific $CD8^+$ T cells in naive and immunized mice as calculated by (**A**) tetramer staining and (**B**) IFN-γ ELISpot assays. (**C**) Correlation between the frequencies of antigen-specific $CD8^+$ T cells in immunized mice as calculated by tetramer staining and IFN-γ ELISpot assays. Average frequencies were used for plotting data. Fitness of curves was determined by the coefficient of determination (r^2).

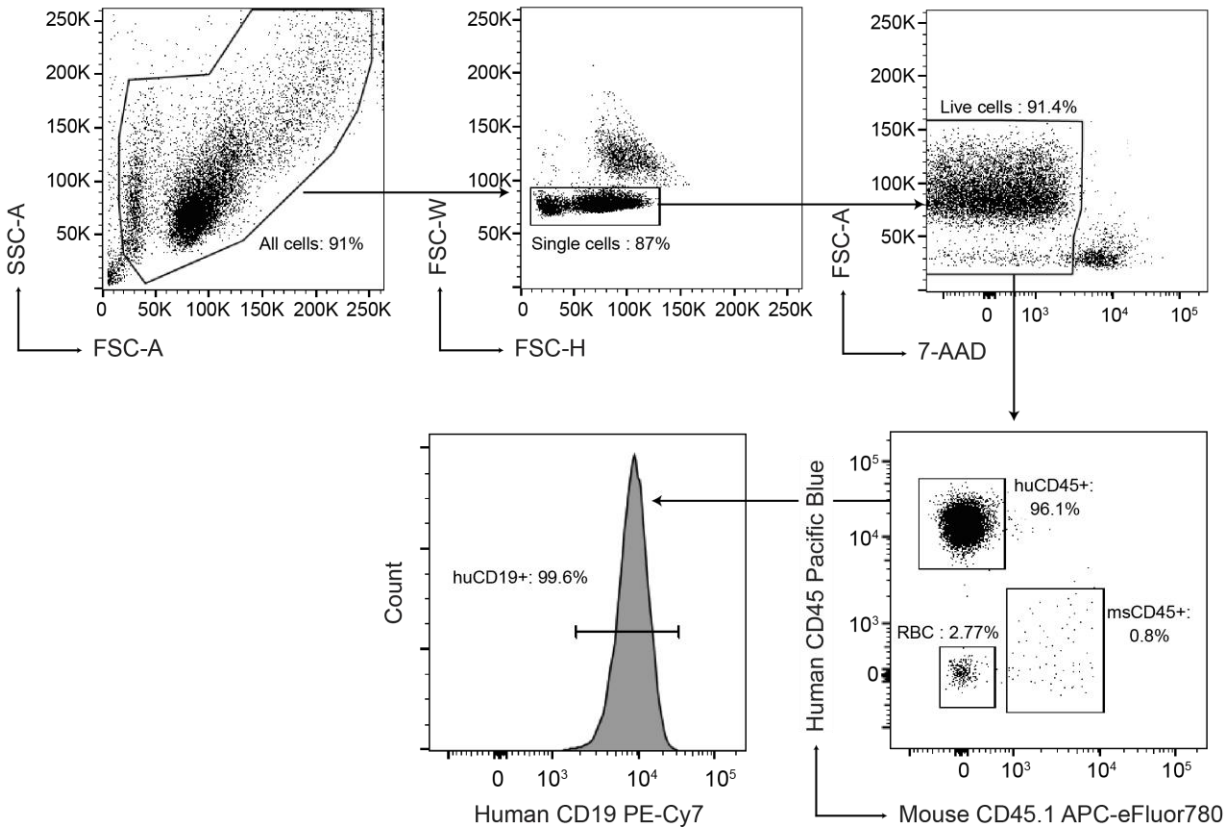


Fig. S8. Purity of the 10H080 B-ALL sample after expansion in NSG mice. After isolation of B-ALL from NSG mice, purity and viability were assessed using flow cytometry. 0.5 x 10^6 cells were stained with anti-human CD45, anti-human CD19, anti-mouse CD45.1 and 7-AAD. Dot plots showing the gating strategy leading to identification of B-ALL cells in one representative sample. B-ALL cells were defined as 7-AAD$^-$huCD45$^+$ cells that homogeneously expressed huCD19 and always represented about 96% of harvested cells. Remaining contaminants after Ficoll gradient were composed of red blood cells (RBC, 2-3%) that do not express MHC at their surface and murine CD45$^+$ cells (about 1%).

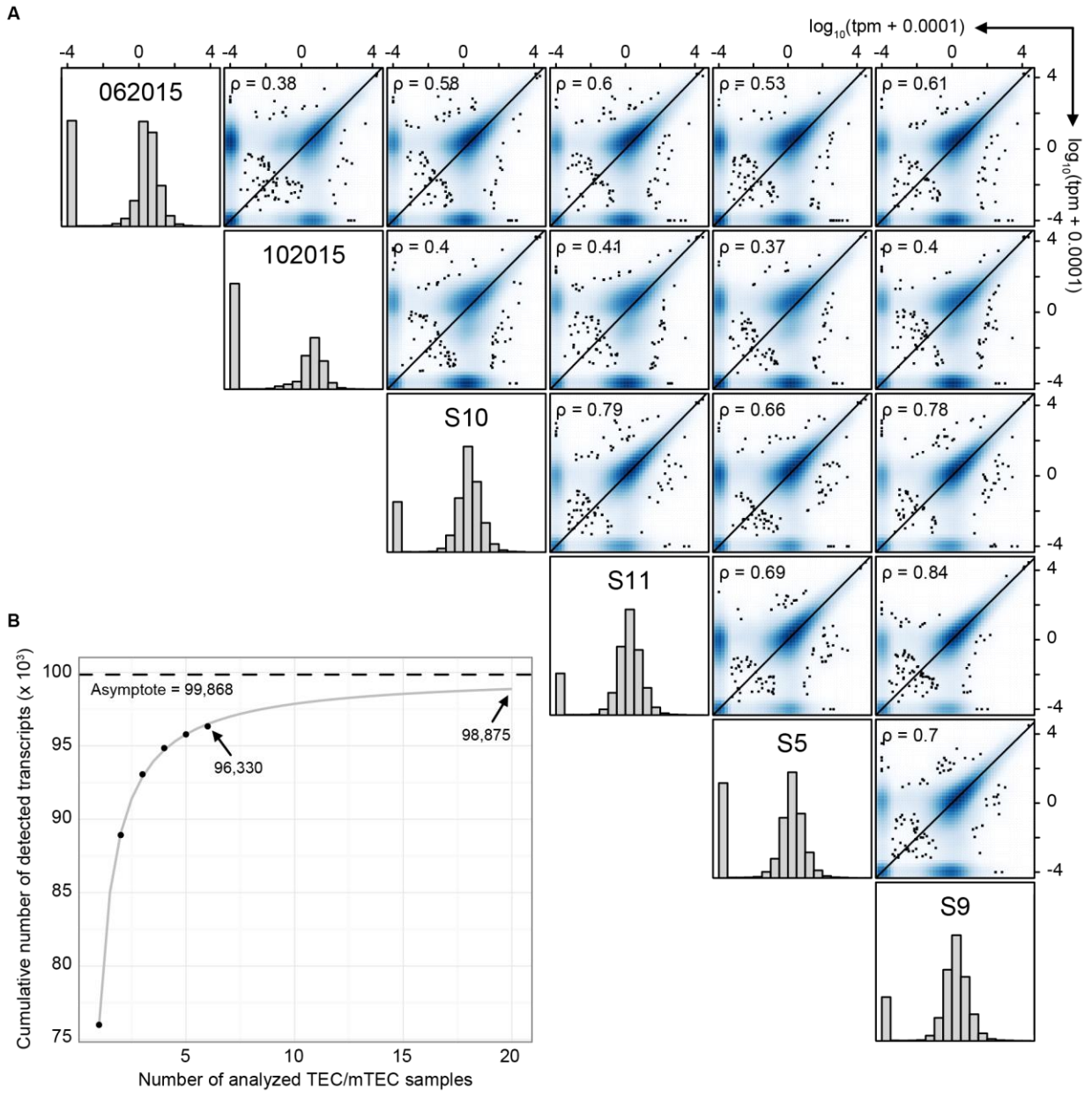


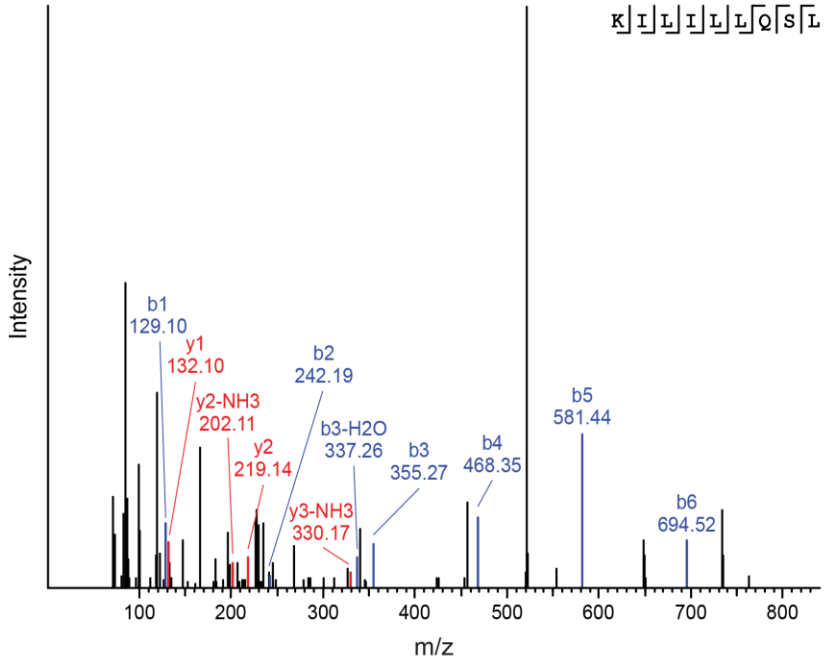
Fig. S9. Overview of the human TEC and mTEC transcriptomic landscapes. (**A**) Human TEC (062015 and 102015) and mTEC (S5 to S11) isolated from unrelated donors display similar transcriptomic profiles. Following RNA-Seq, we selected transcripts expressed in at least one donor with a tpm > 1, as estimated by kallisto, to plot all one-to-one scatter plots. The Spearman's rank correlation coefficient (ρ) is indicated at the top left corner of each graph and the black line represents identical expression of transcripts. (**B**) RNA-Seq of additional human

TEC/mTEC samples should result in a minimal gain of information. Using our set of expressed transcripts (tpm > 1 in at least one sample), we extrapolated the cumulative number of transcripts (*cT*) that should be detected by adding additional samples to our cohorts (*nS*, see section **Cumulative number of transcripts detected in TEC and mTEC samples** of the **Materials and Methods** section) using the following function: $cT = \frac{a \times (nS-1)}{[b+ (nS-1)]} + c$ with a = 23,892.73, b = 0.8243389 and c = 75,976.11 (grey line). On the graph, we indicated the cumulative number of transcripts detected by analyzing nS = 6 (our cohort, black dots) or nS = 20 samples, as well as the total number of transcripts that should be detected, which corresponds to $\lim_{nS\to\infty} \left(\frac{a \times (nS-1)}{[b+ (nS-1)]} + c\right) = a + c = 99{,}868$ (asymptote value).

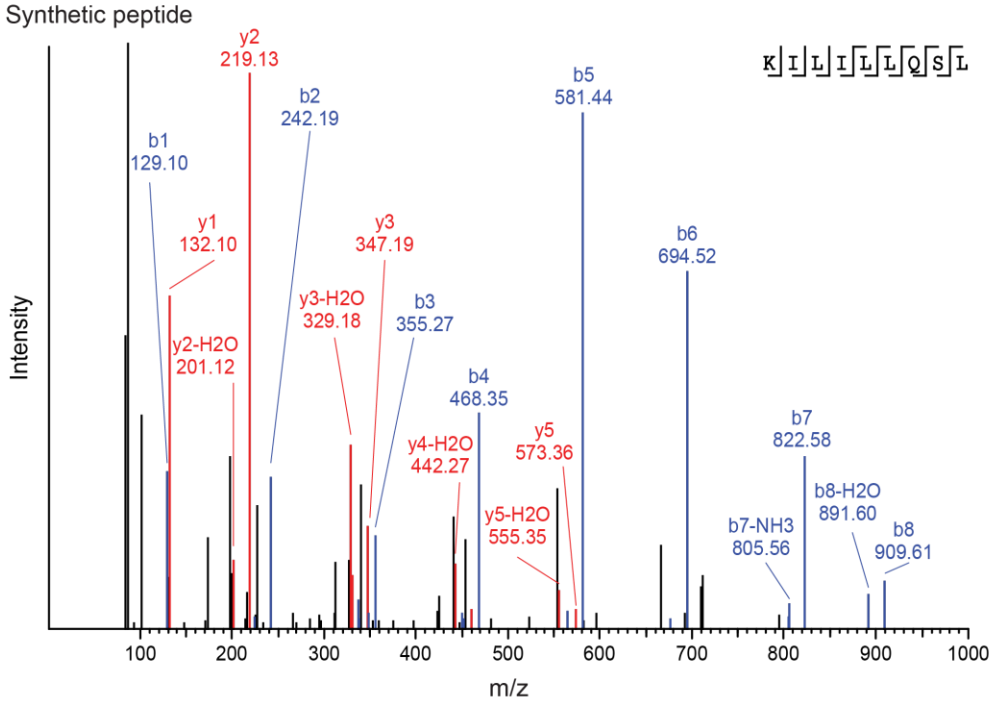
A

KILILLQSL - ERE aeTSA

Endogenous peptide



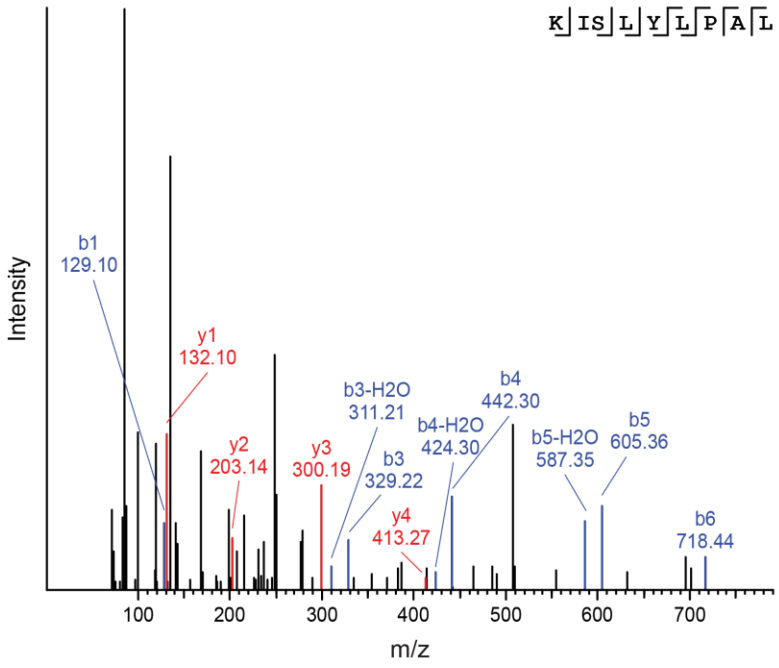
Synthetic peptide



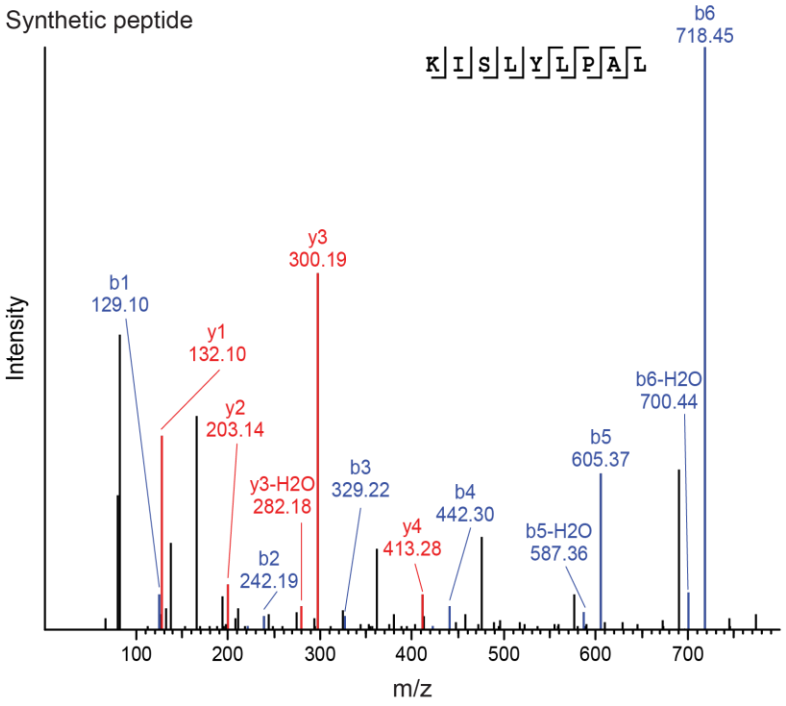
B

KISLYLPAL - ERE aeTSA

Endogenous peptide



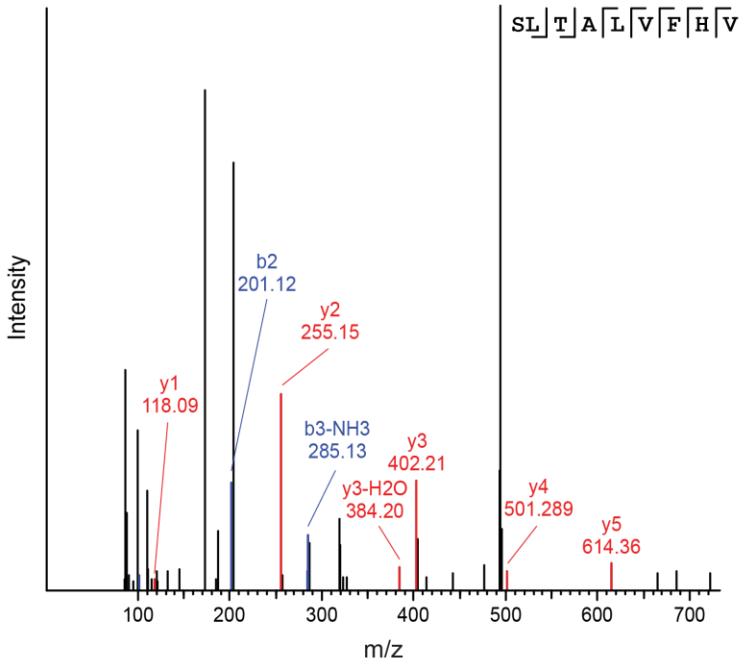
Synthetic peptide



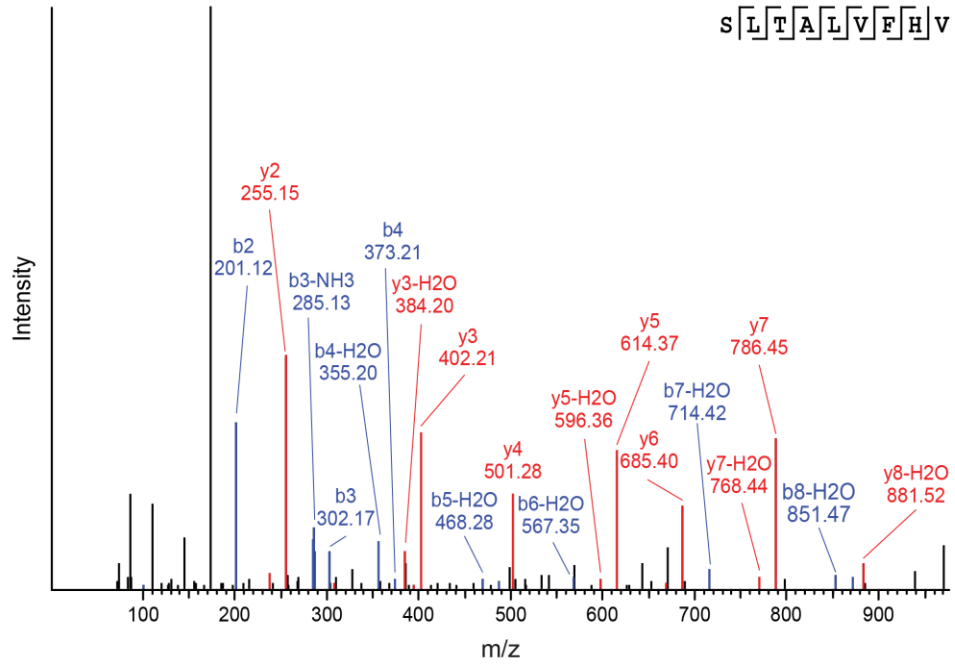
C

SLTALVFHV - aeTSA

Endogenous peptide



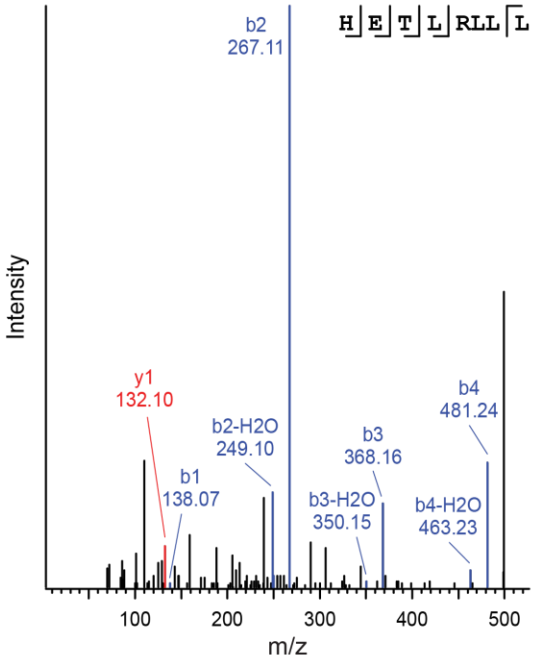
Synthetic peptide



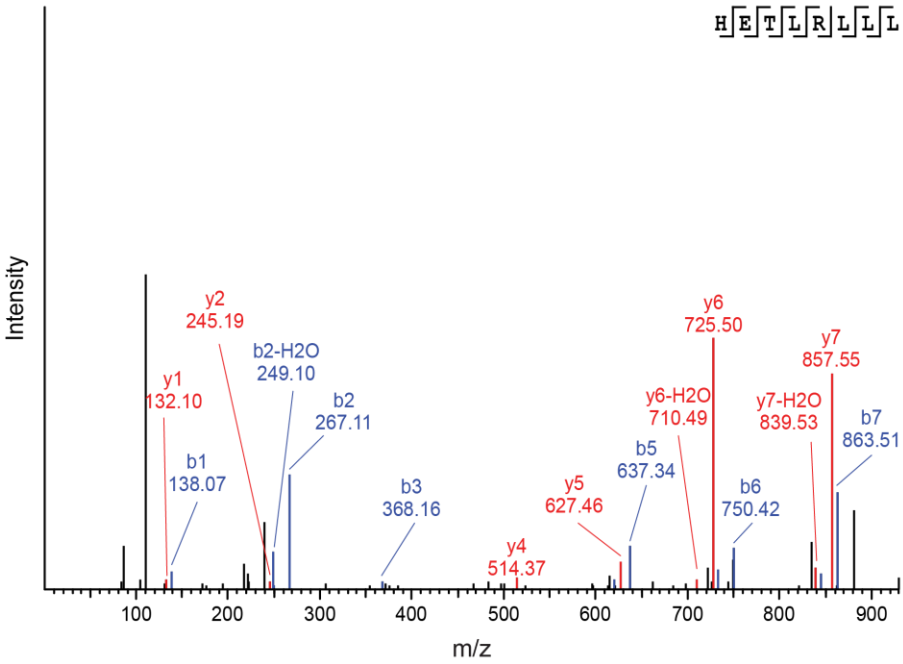
D

HETLRLLL - aeTSA

Endogenous peptide



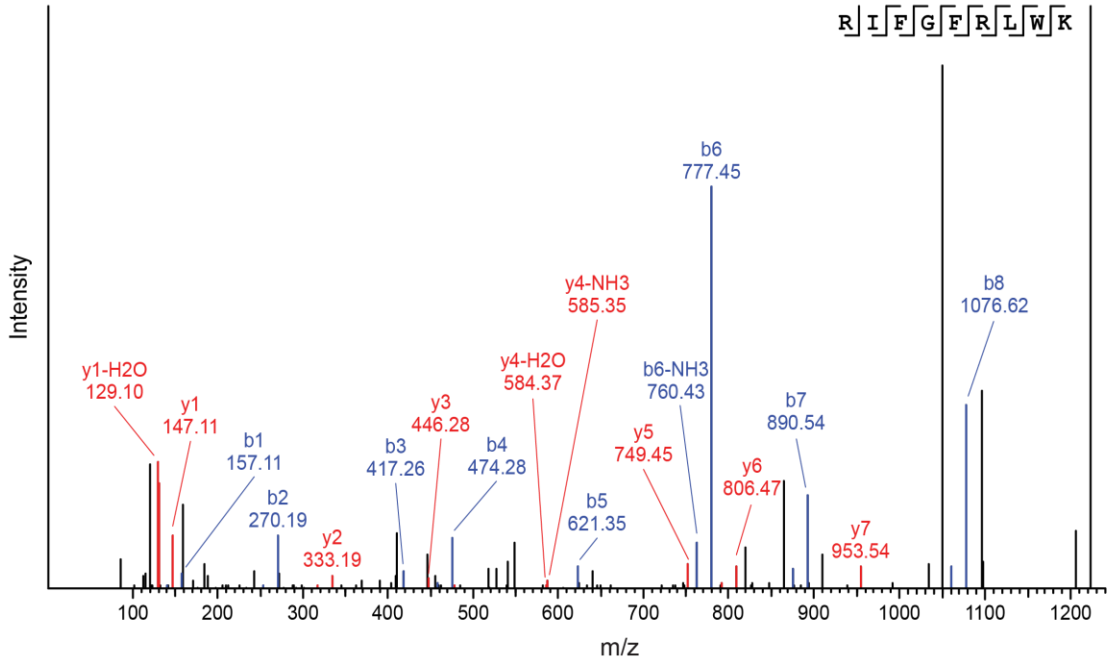
Synthetic peptide



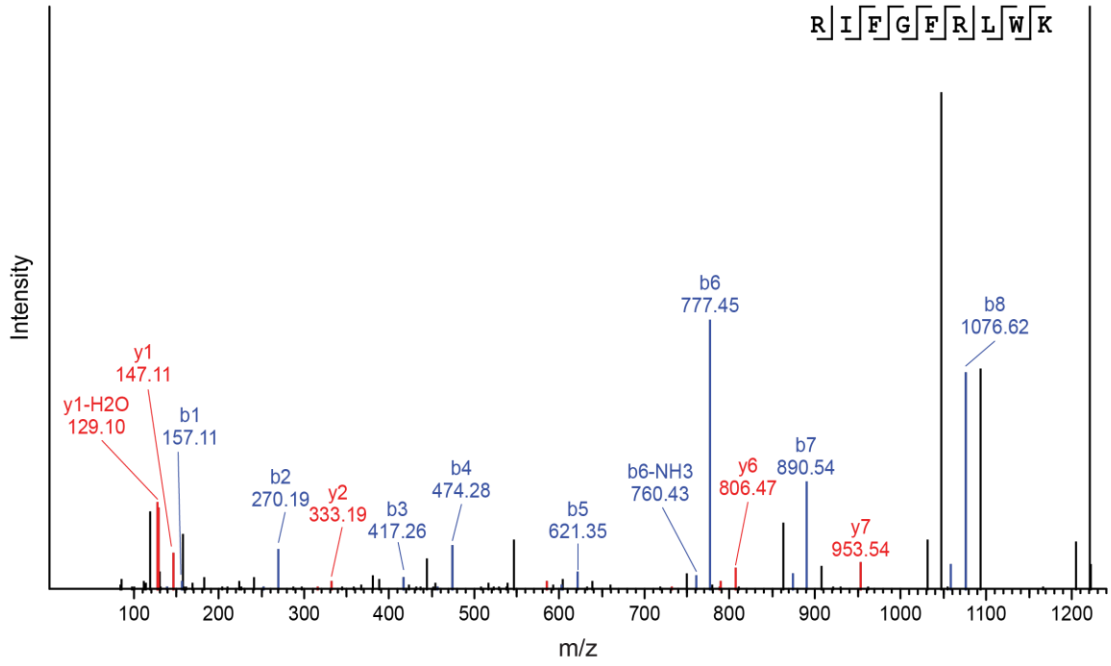
E

RIFGFRLWK - aeTSA

Endogenous peptide



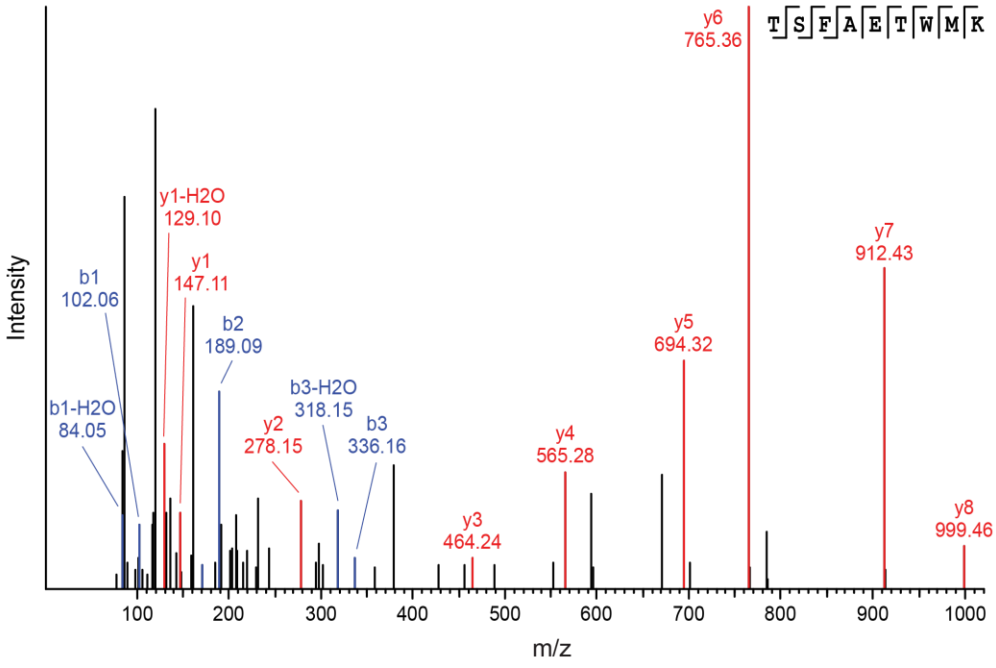
Synthetic peptide



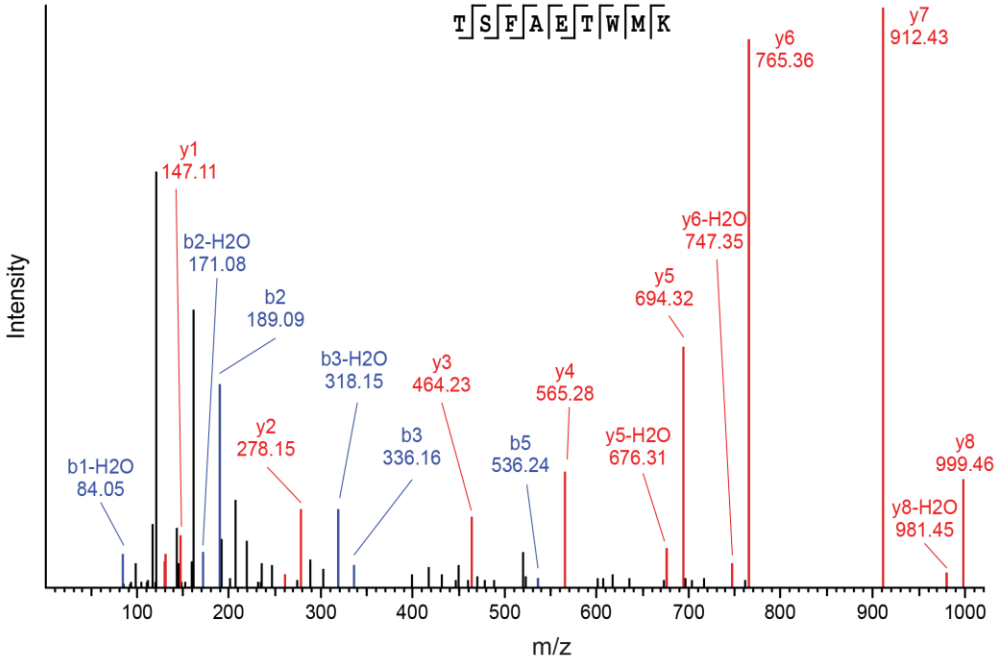
F

TSFAETWMK - ERE aeTSA

Endogenous peptide



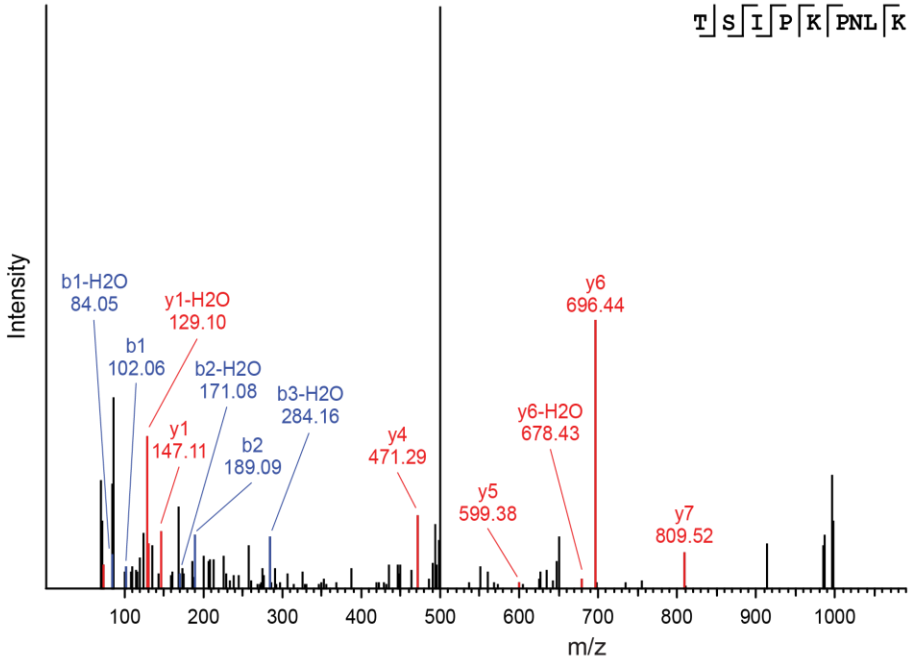
Synthetic peptide



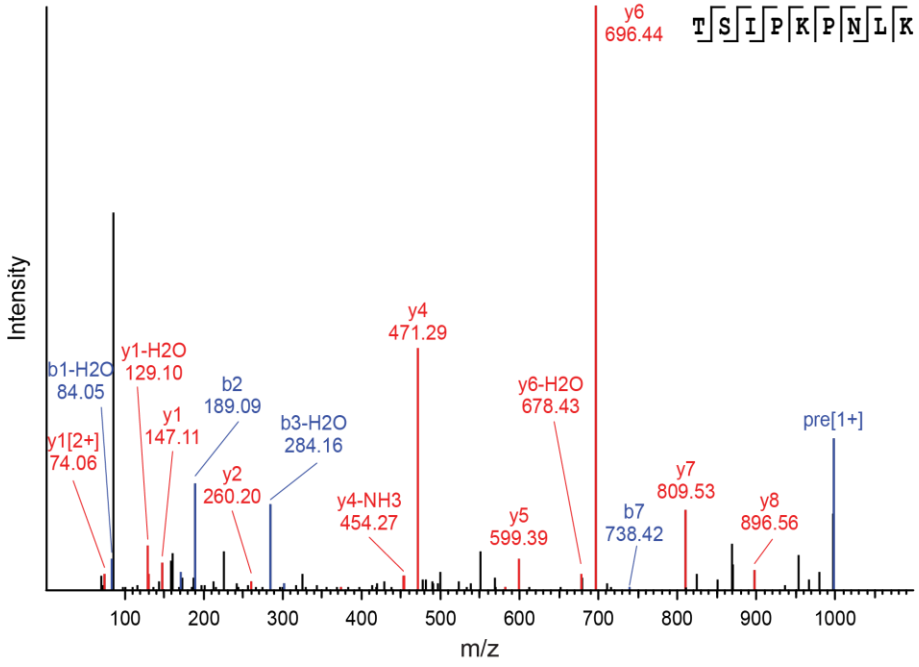
G

TSIPKPNLK - aeTSA

Endogenous peptide



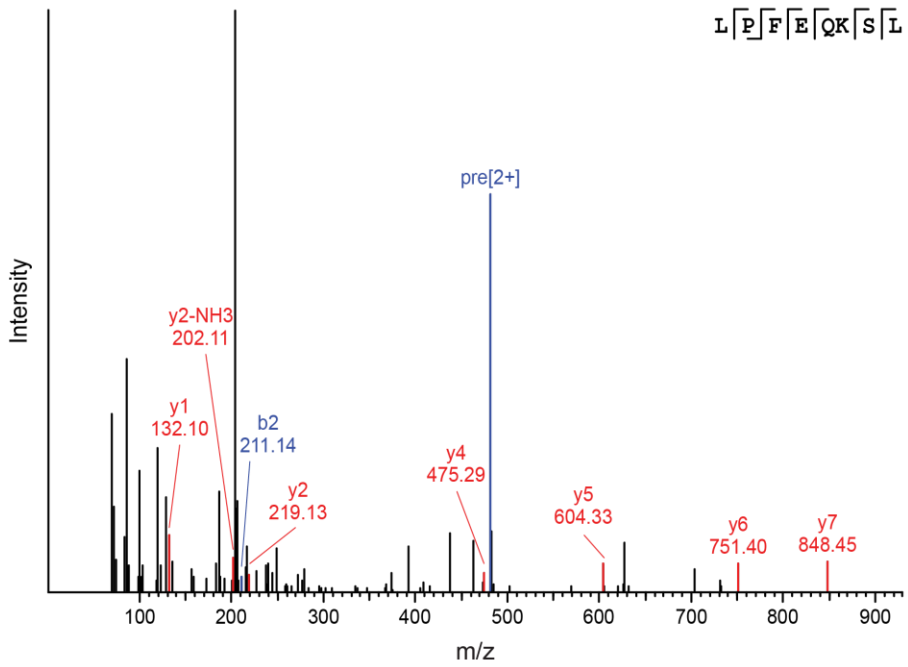
Synthetic peptide



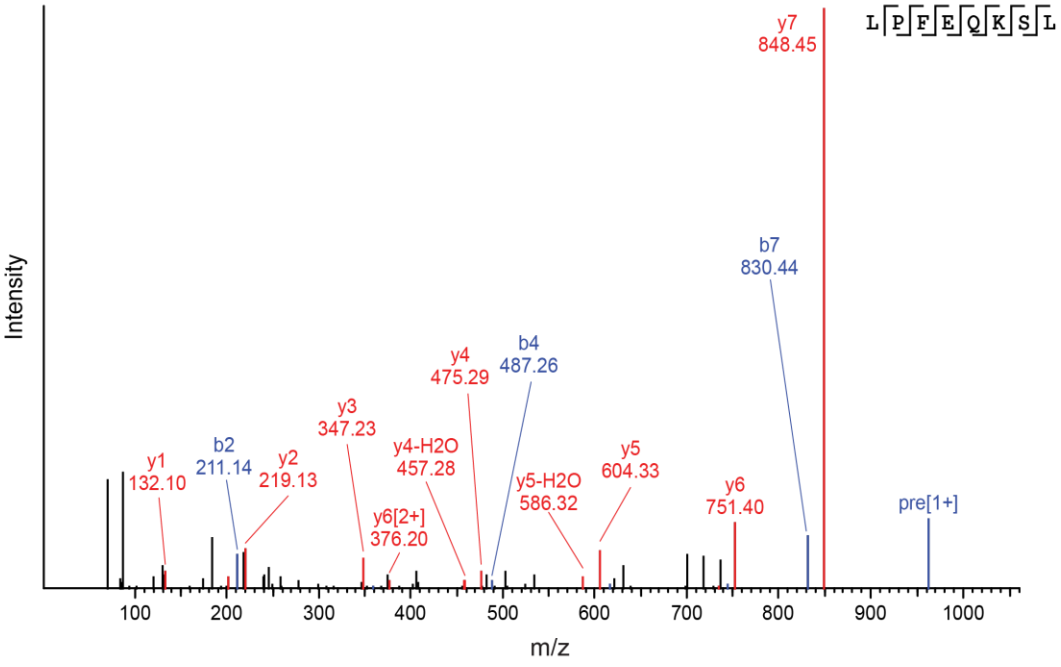
H

LPFEQKSL - aeTSA

Endogenous peptide



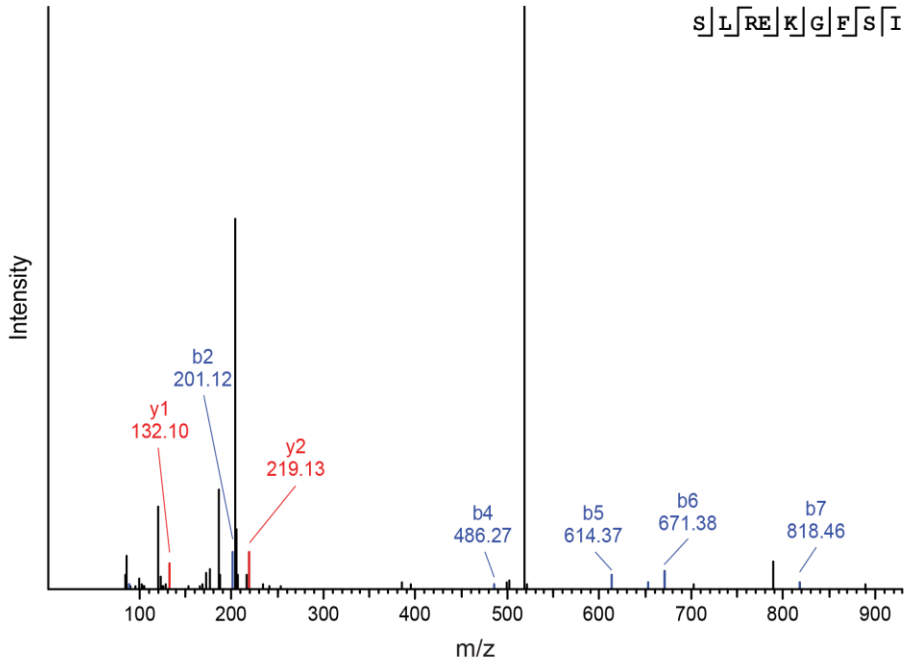
Synthetic peptide



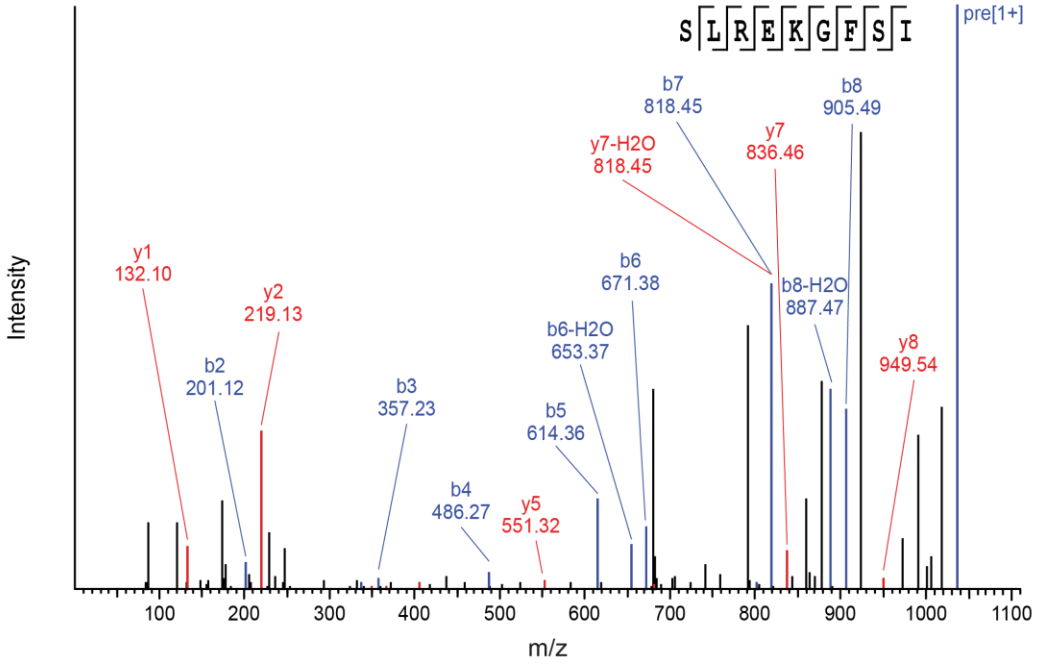
I

SLREKGFSI - aeTSA

Endogenous peptide



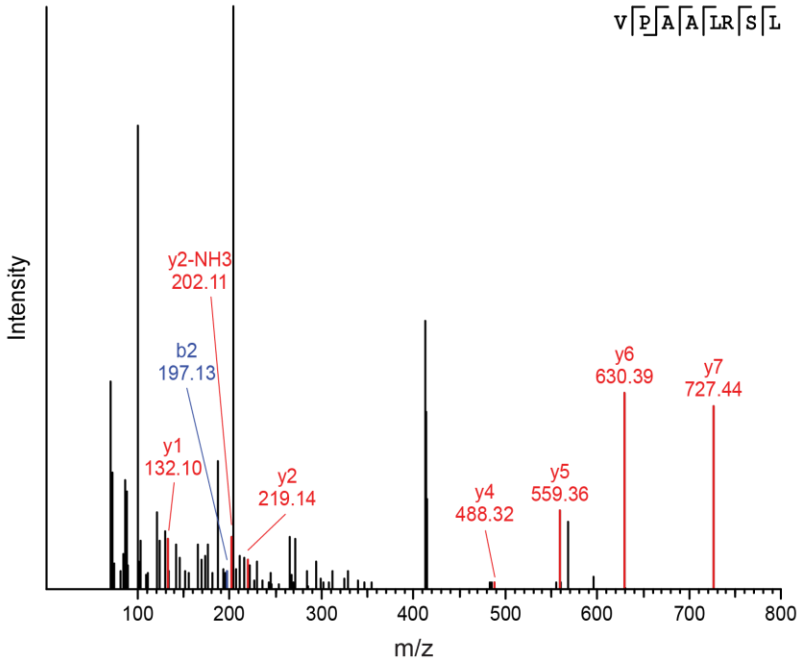
Synthetic peptide



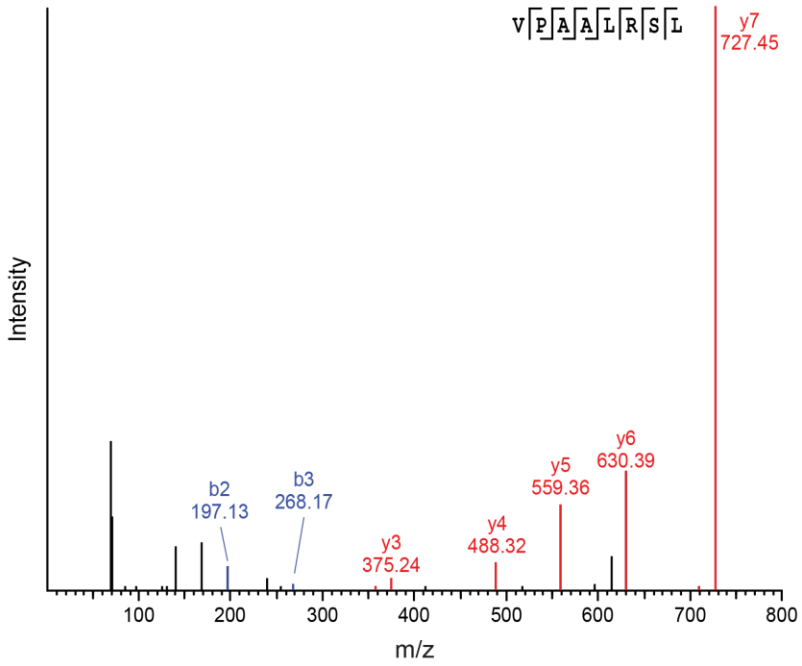
J

VPAALRSL - aeTSA

Endogenous peptide



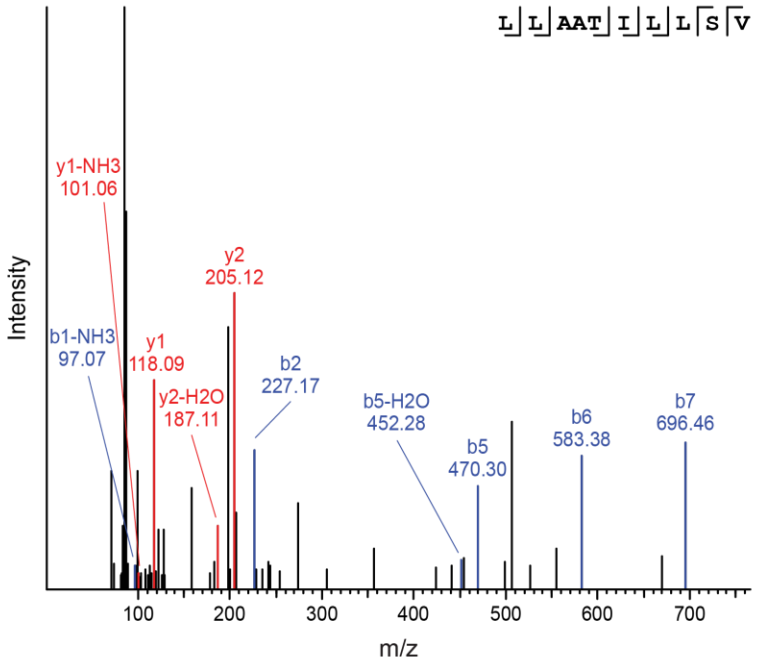
Synthetic peptide



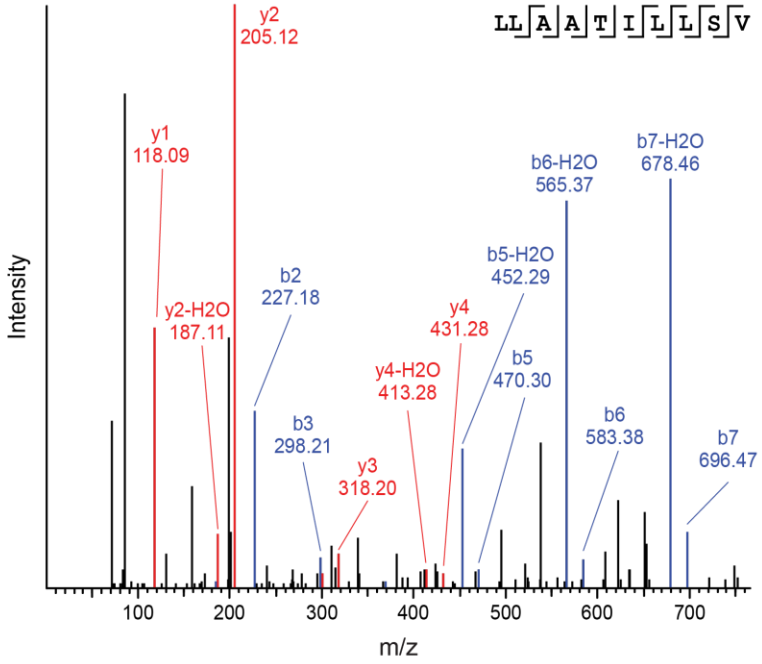
K

LLAATILLSV - aeTSA

Endogenous peptide



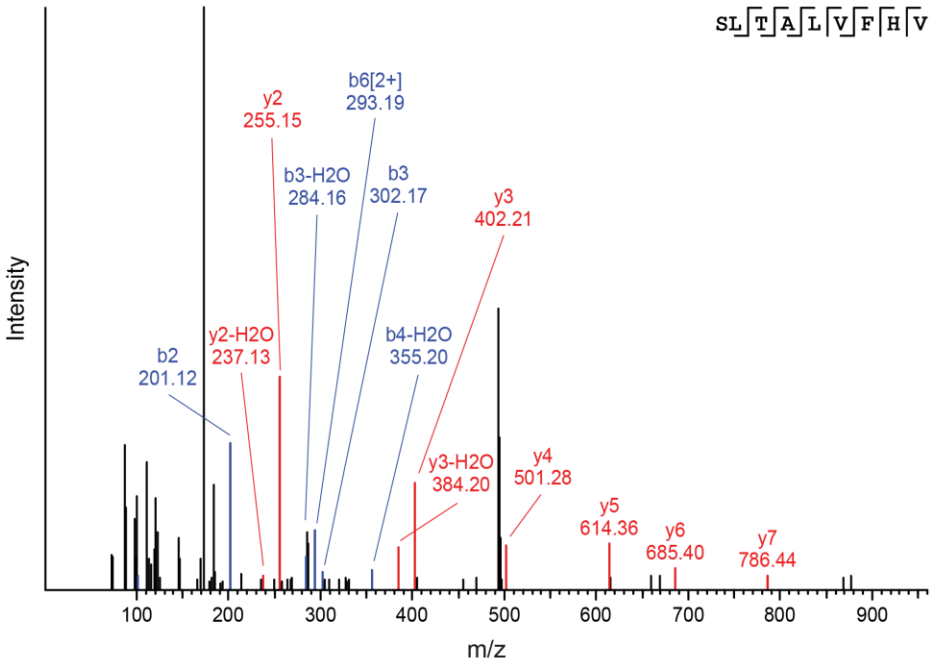
Synthetic peptide



L

SLTALVFHV - aeTSA

Endogenous peptide



Synthetic peptide

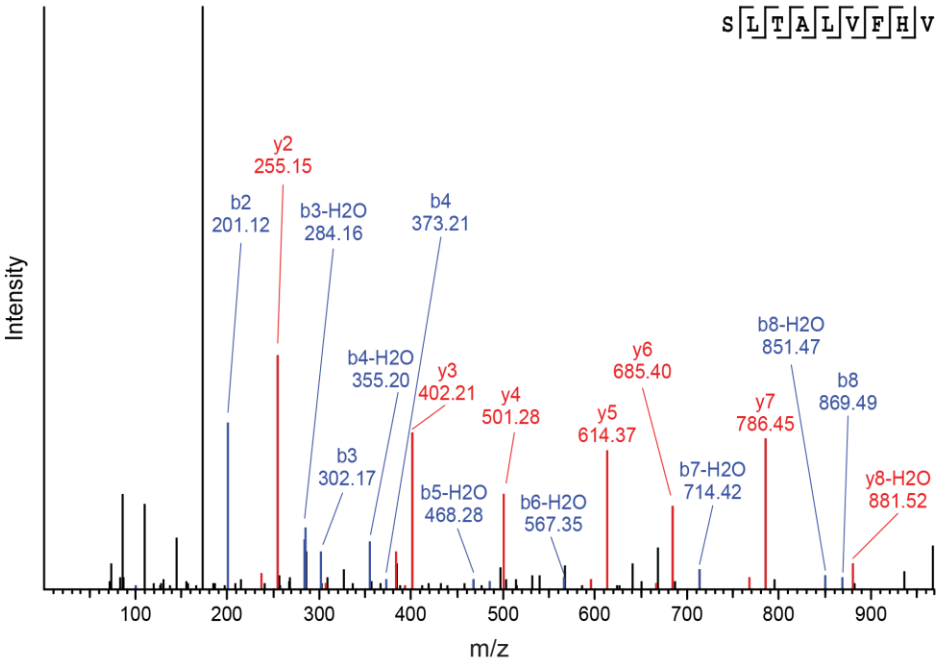
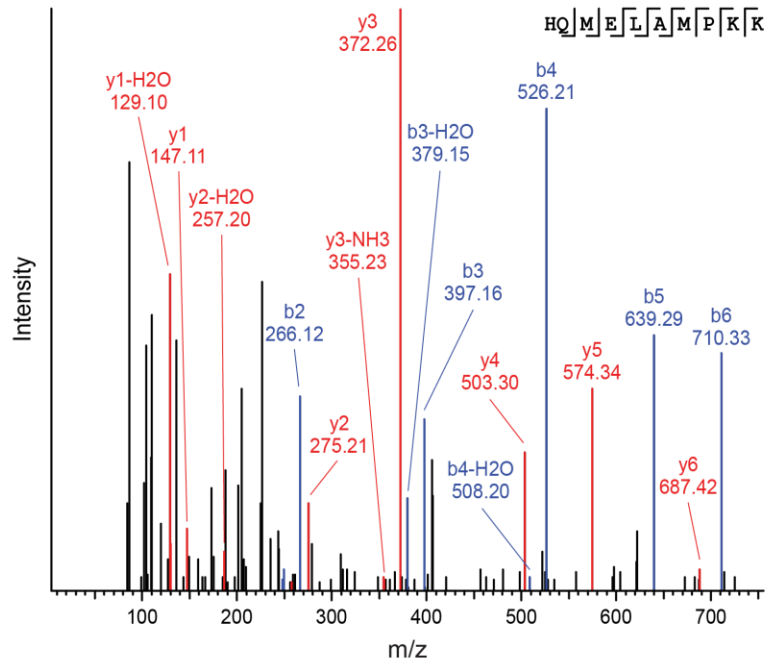


Fig. S10. MS validation of B-ALL TSA candidates using synthetic analogs. Synthetic and endogenous MS/MS spectra for TSA candidates identified in each of our four B-ALL specimens: (**A-C**) 07H103, (**D-G**) 10H080, (**H-J**) 10H118 and (**K-L**) 12H018. See section **MS validation of TSA candidates** of the **Supplementary Materials** for details.

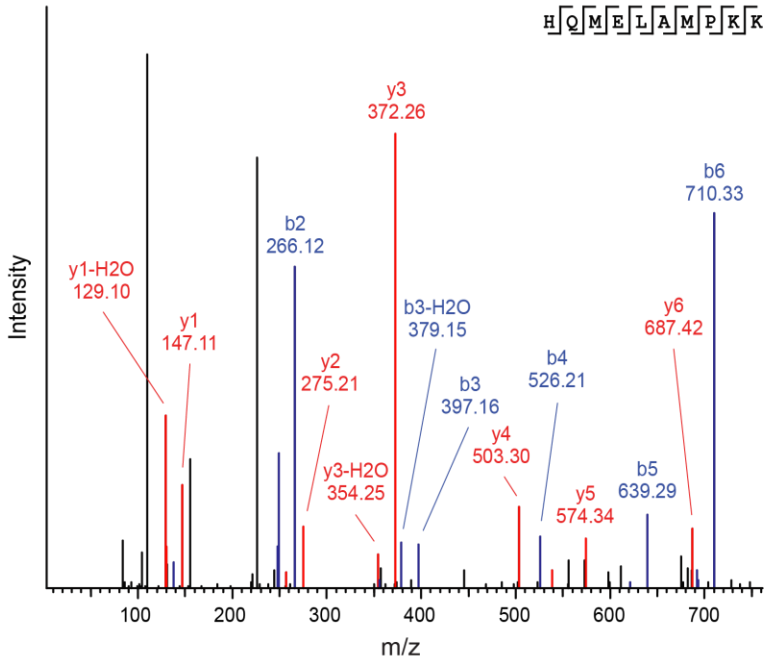
A

HQMELAMPKK - aeTSA

Endogenous peptide



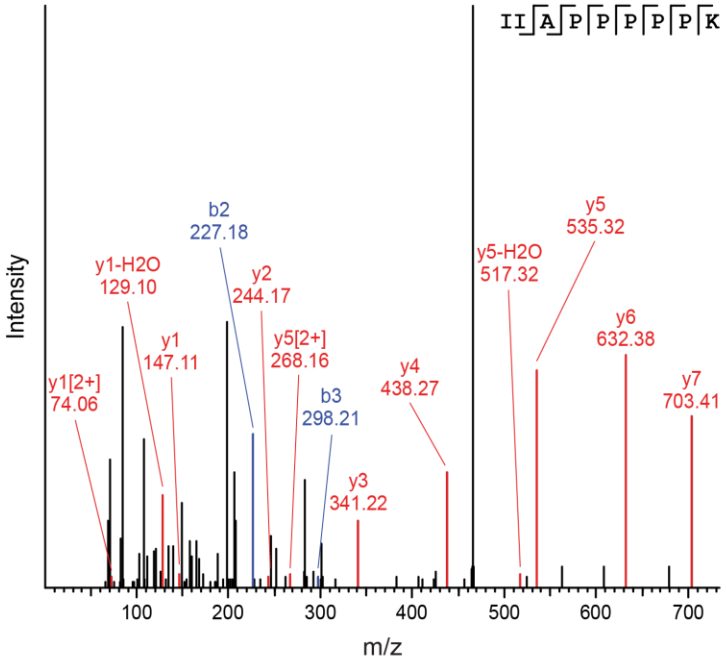
Synthetic peptide



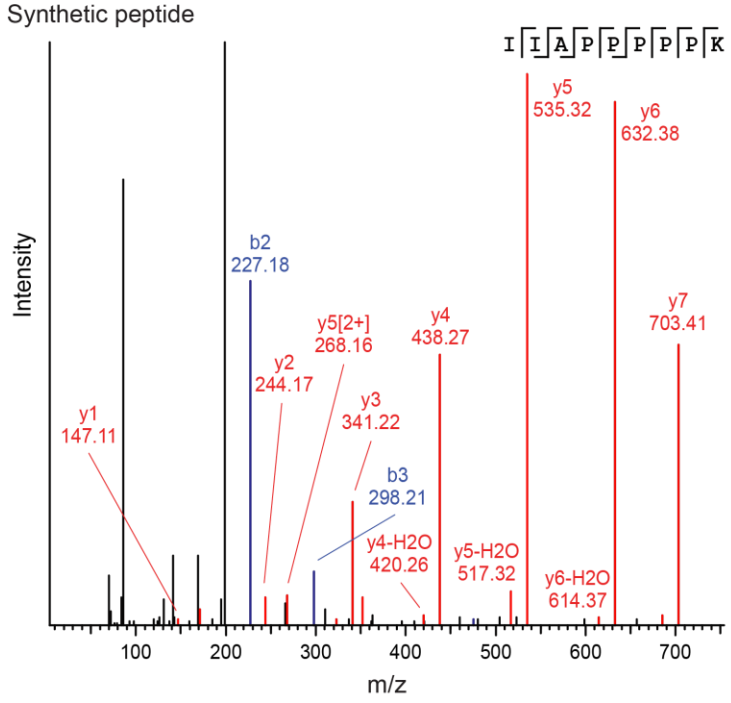
B

IIAPPPPPK - aeTSA

Endogenous peptide



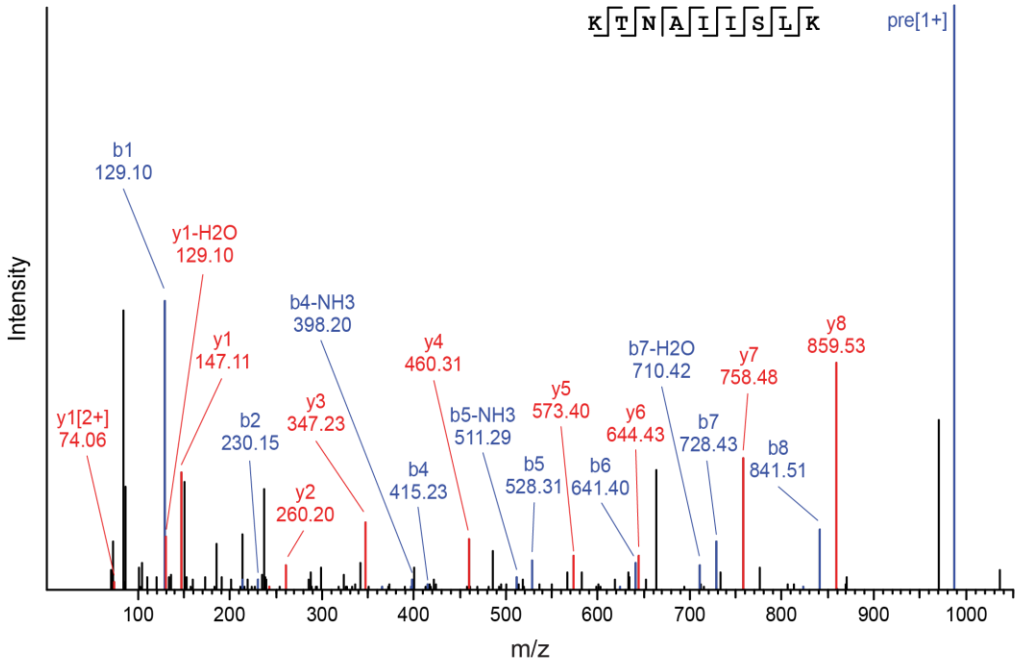
Synthetic peptide



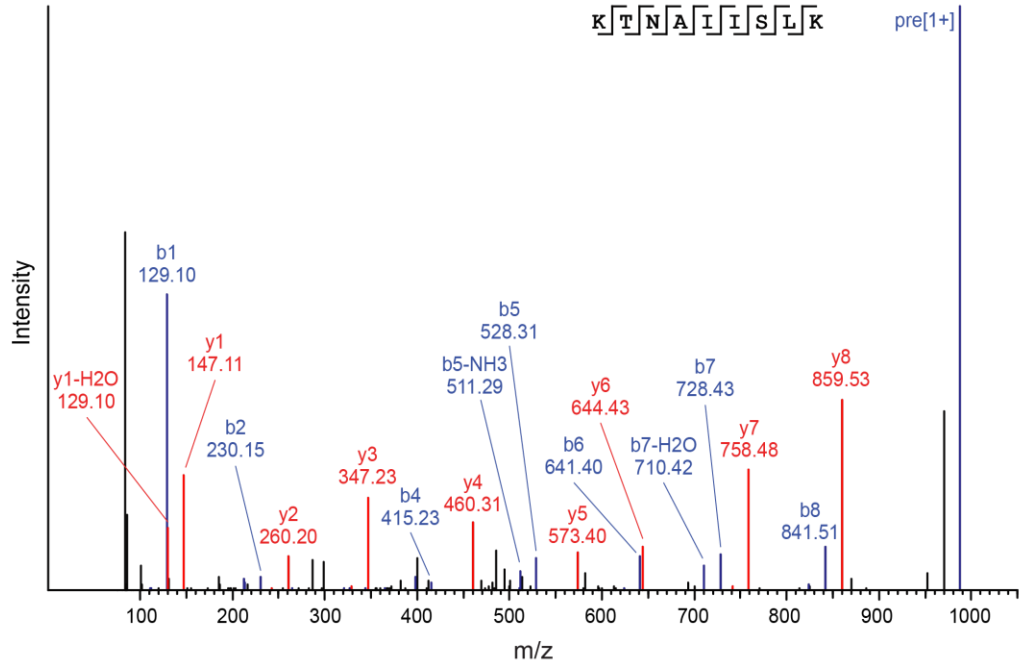
C

KTNAIISLK - aeTSA

Endogenous peptide



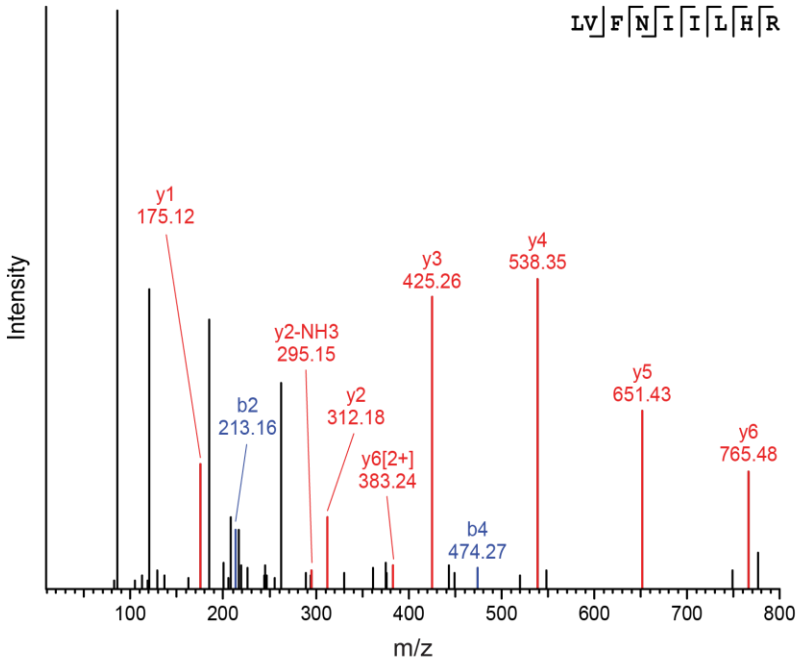
Synthetic peptide



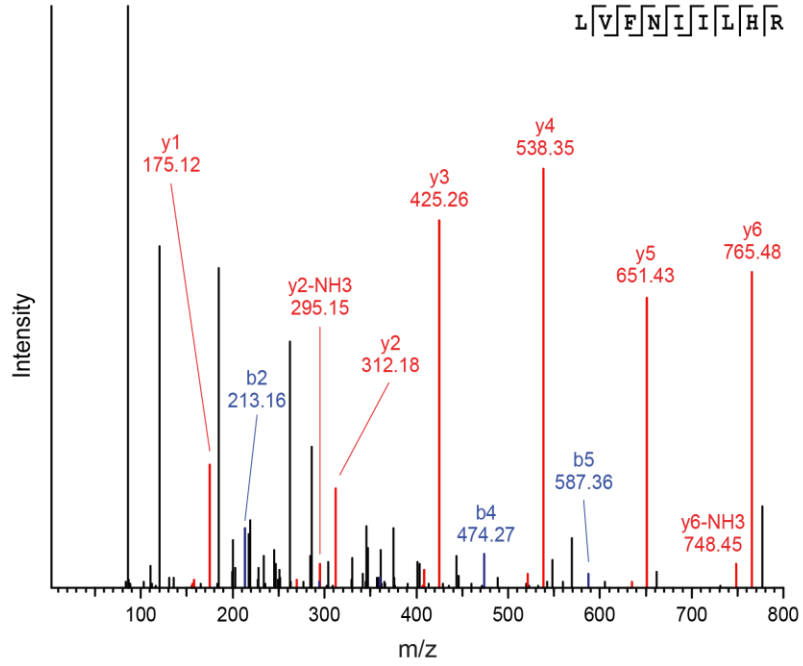
D

LVFNIILHR - aeTSA

Endogenous peptide



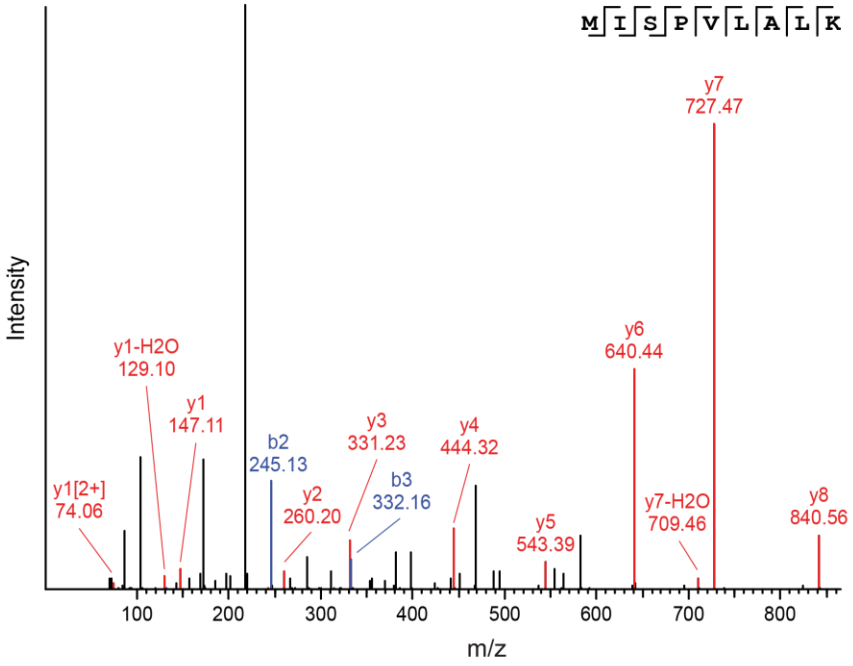
Synthetic peptide



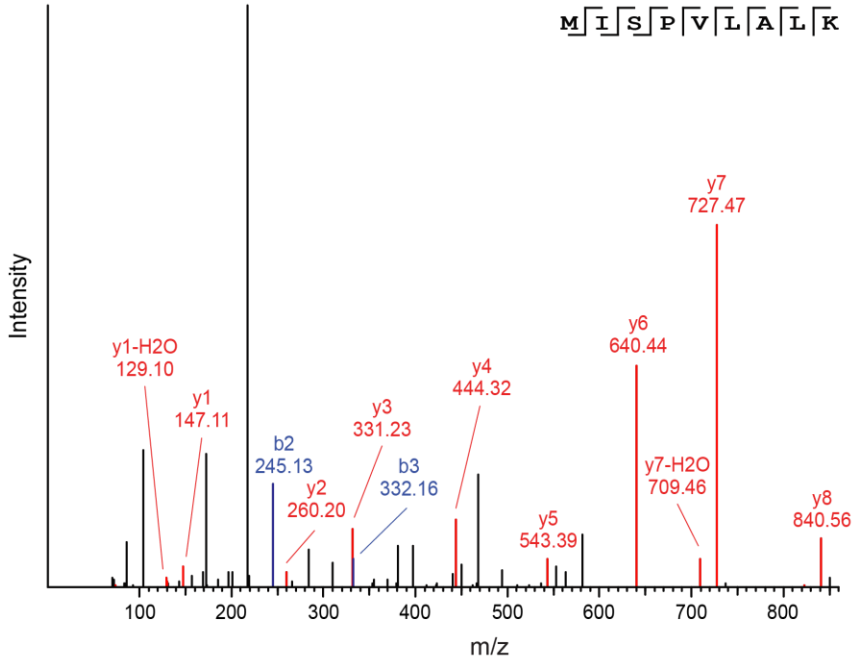
E

MISPVLALK - aeTSA

Endogenous peptide



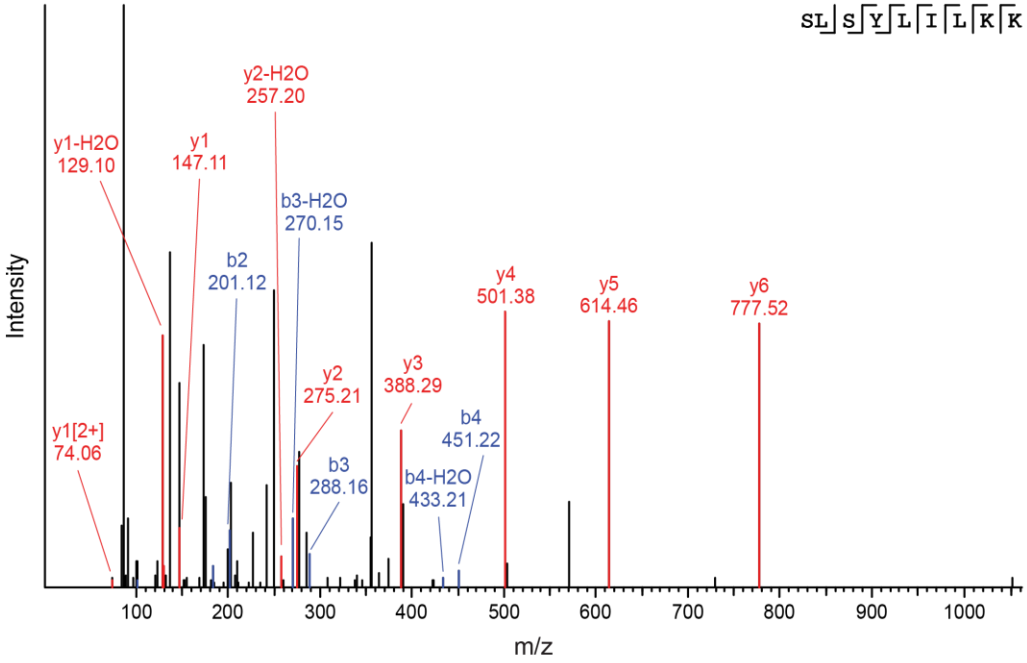
Synthetic peptide



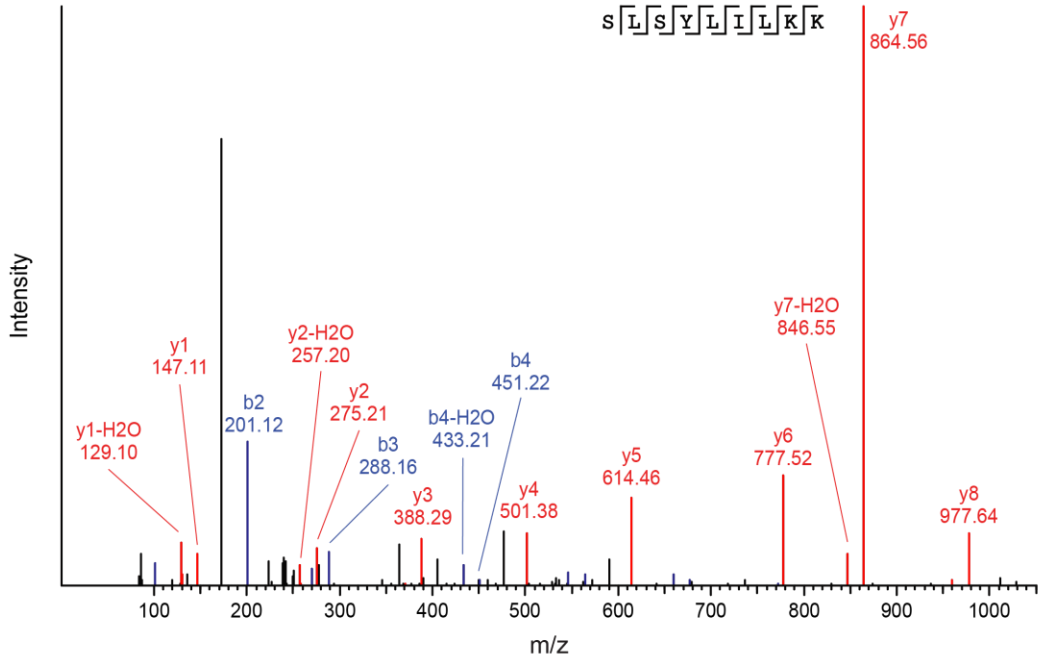
F

SLSYLILKK - aeTSA

Endogenous peptide

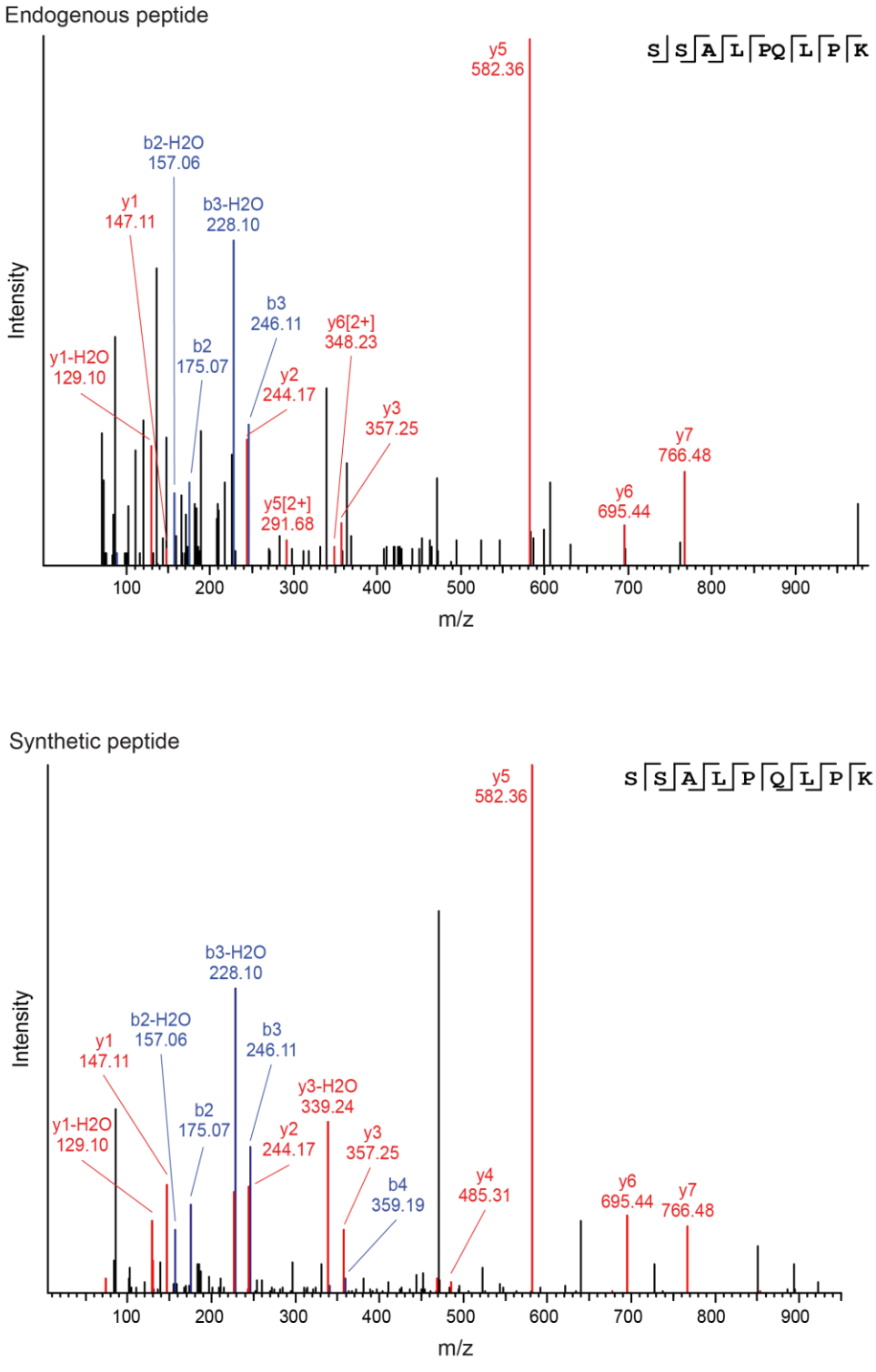


Synthetic peptide



G

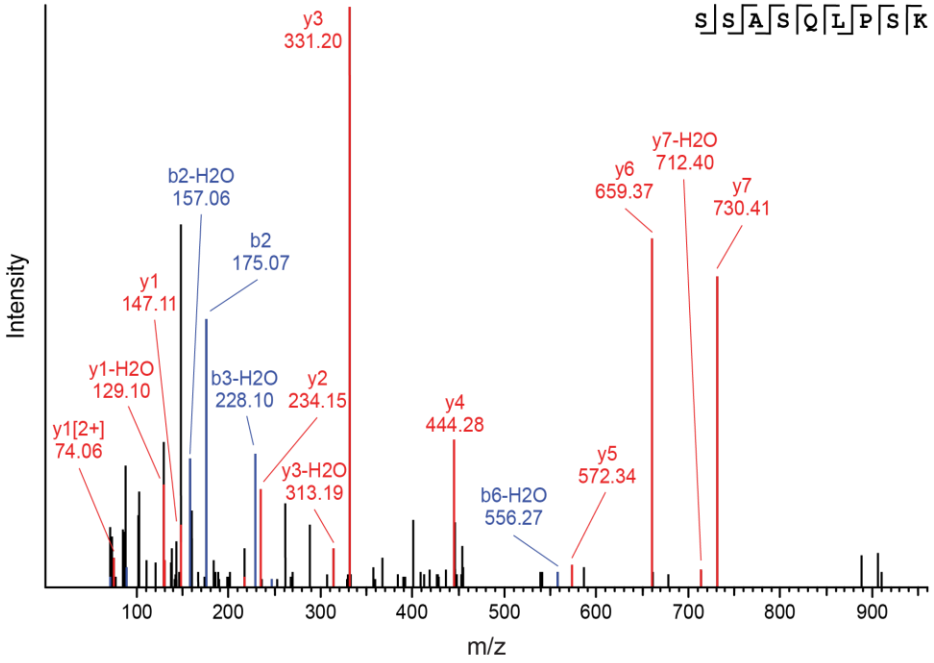
SSALPQLPK - aeTSA



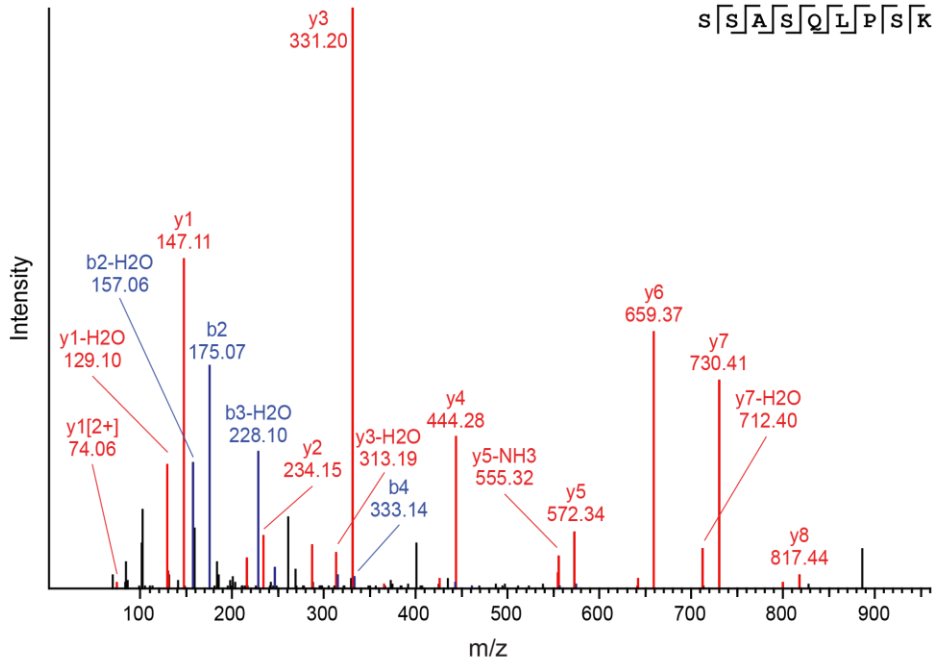
H

SSASQLPSK - ERE aeTSA

Endogenous peptide



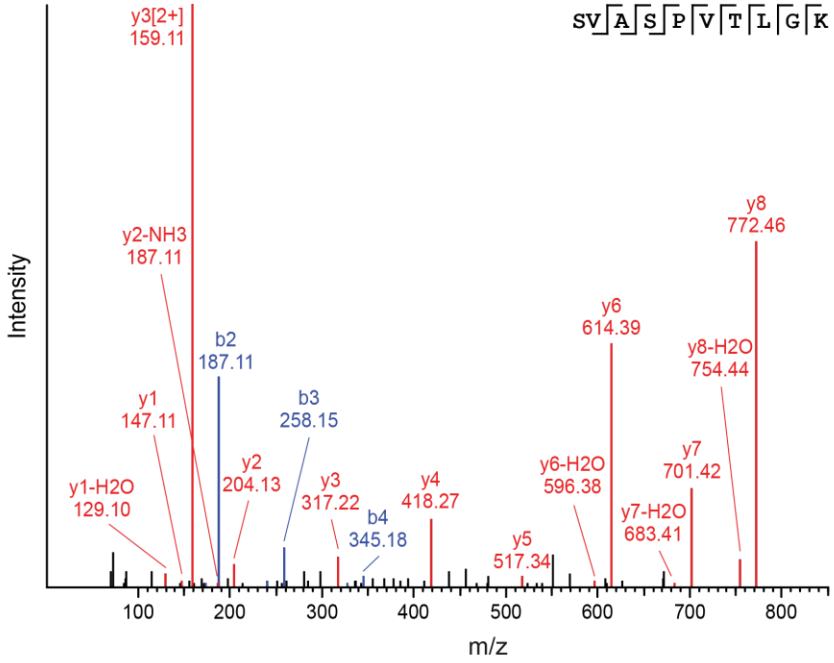
Synthetic peptide



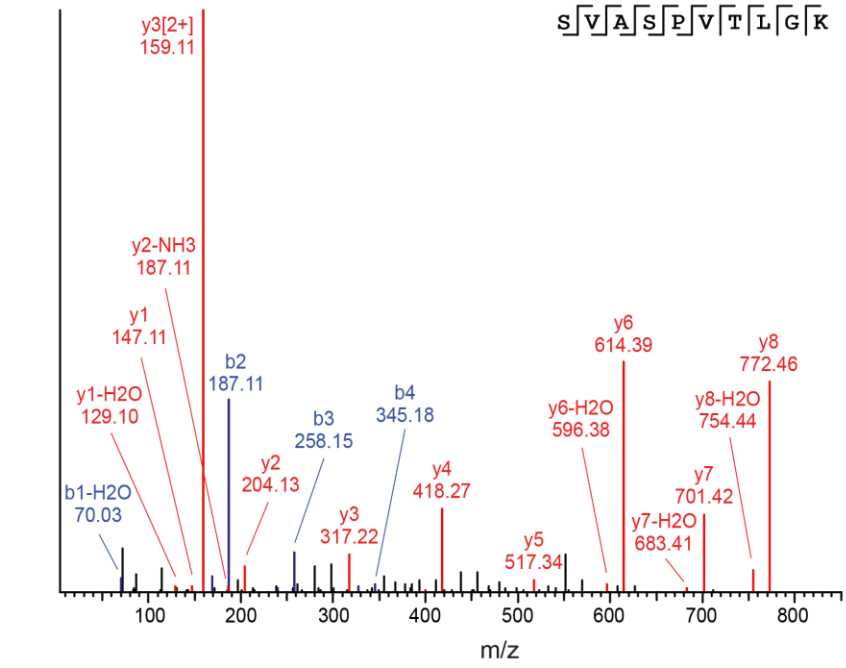
I

SVASPVTLGK - aeTSA

Endogenous peptide



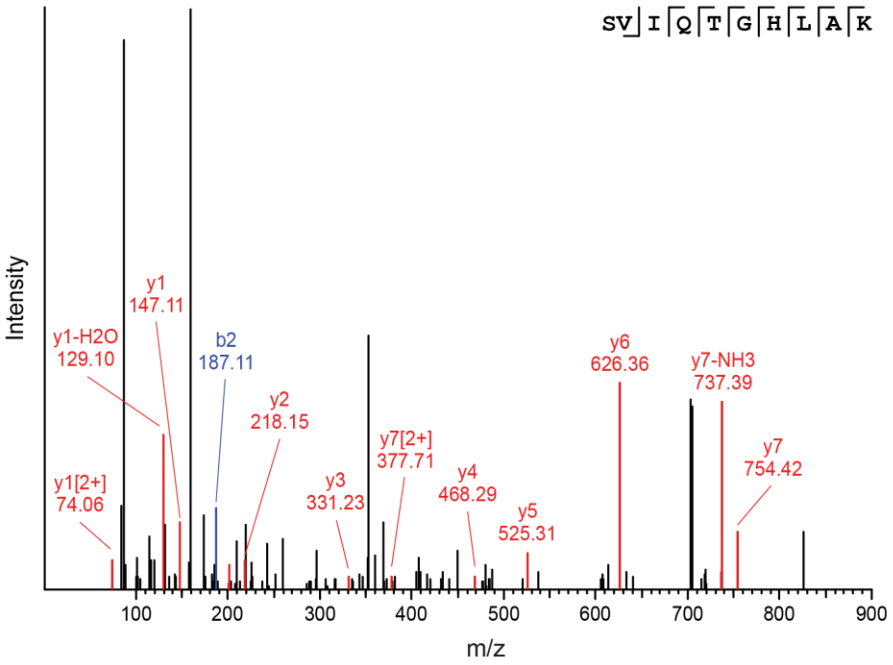
Synthetic peptide



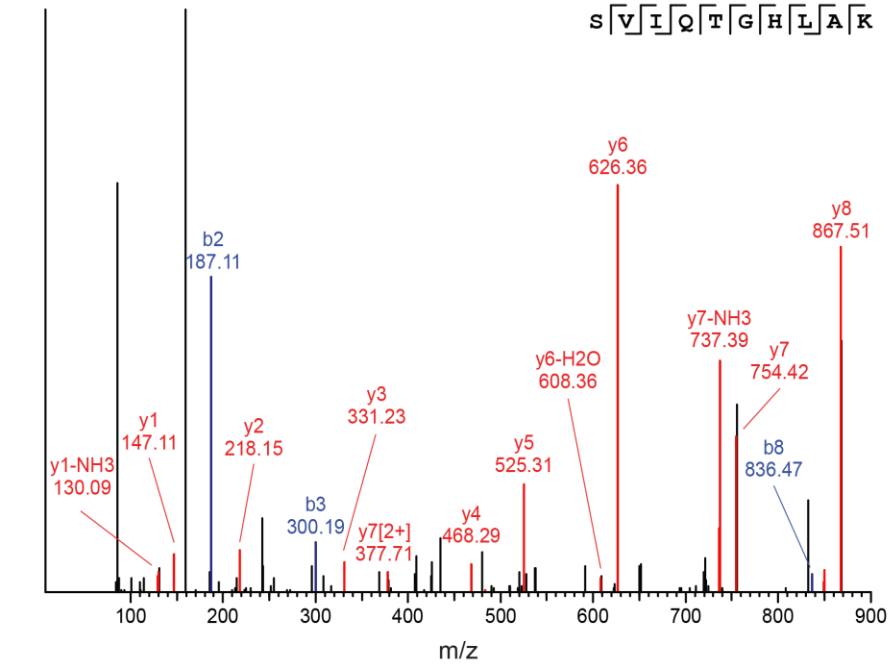
J

SVIQTGHLAK - aeTSA

Endogenous peptide



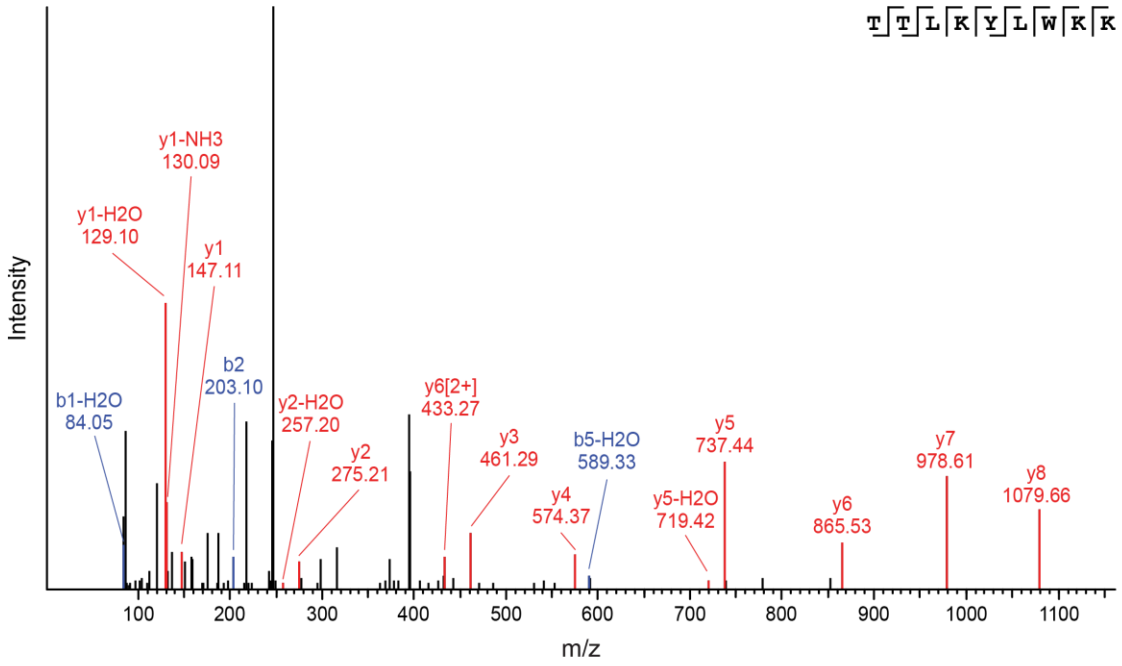
Synthetic peptide



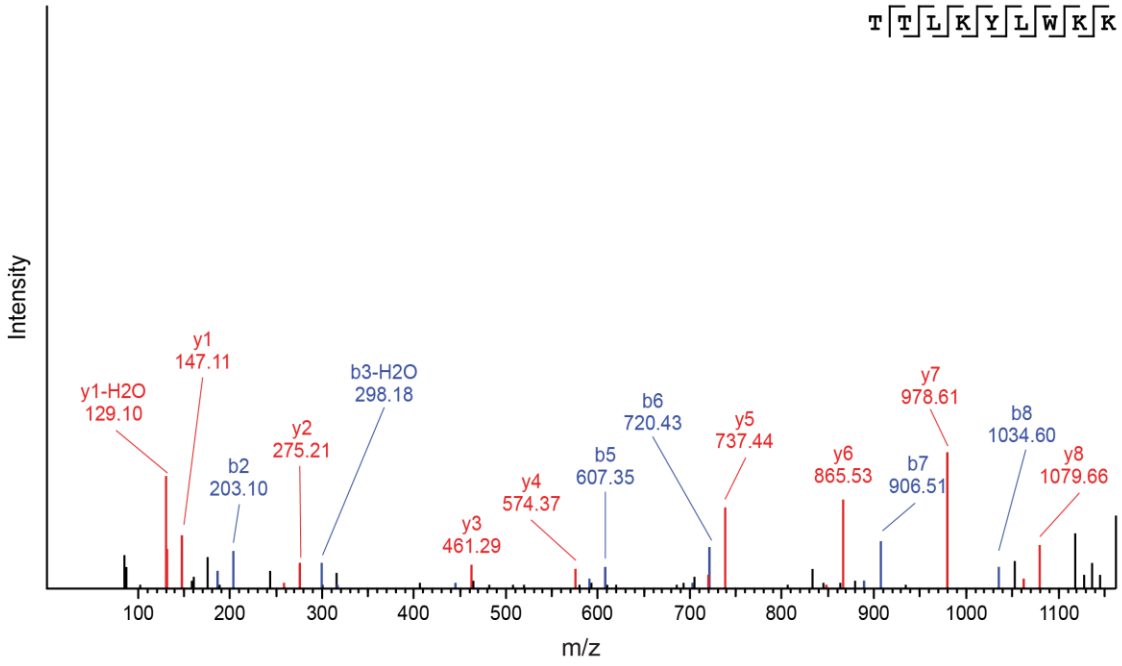
K

TTLKYLWKK - aeTSA

Endogenous peptide



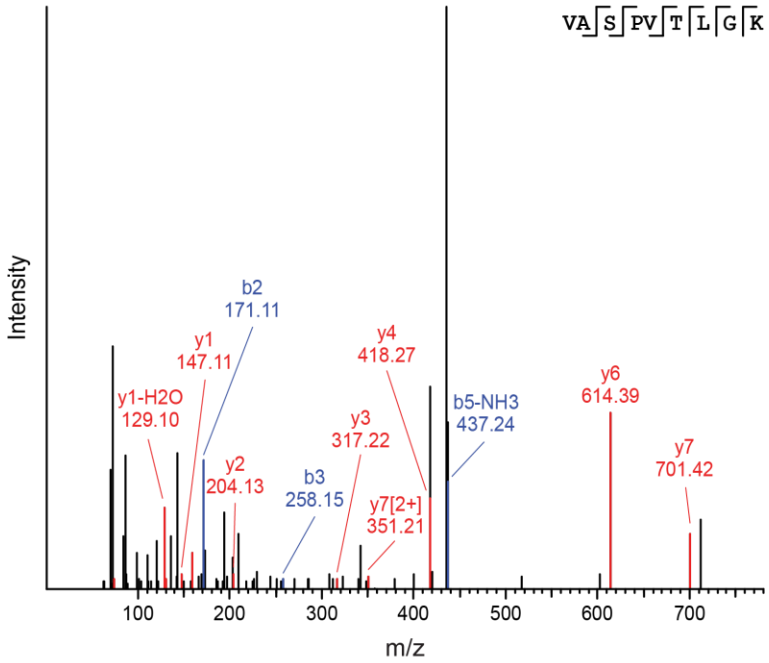
Synthetic peptide



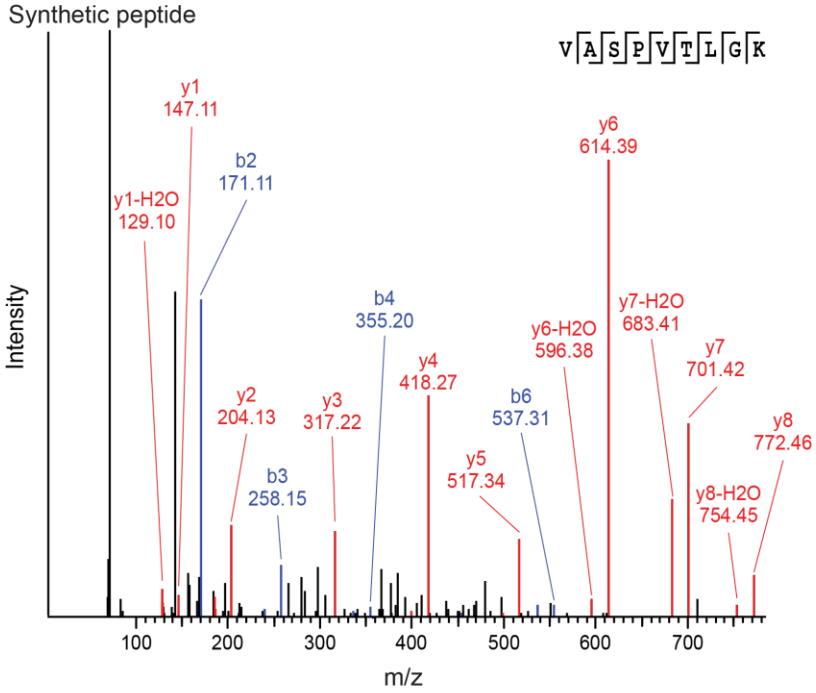
L

VASPVTLGK - aeTSA

Endogenous peptide



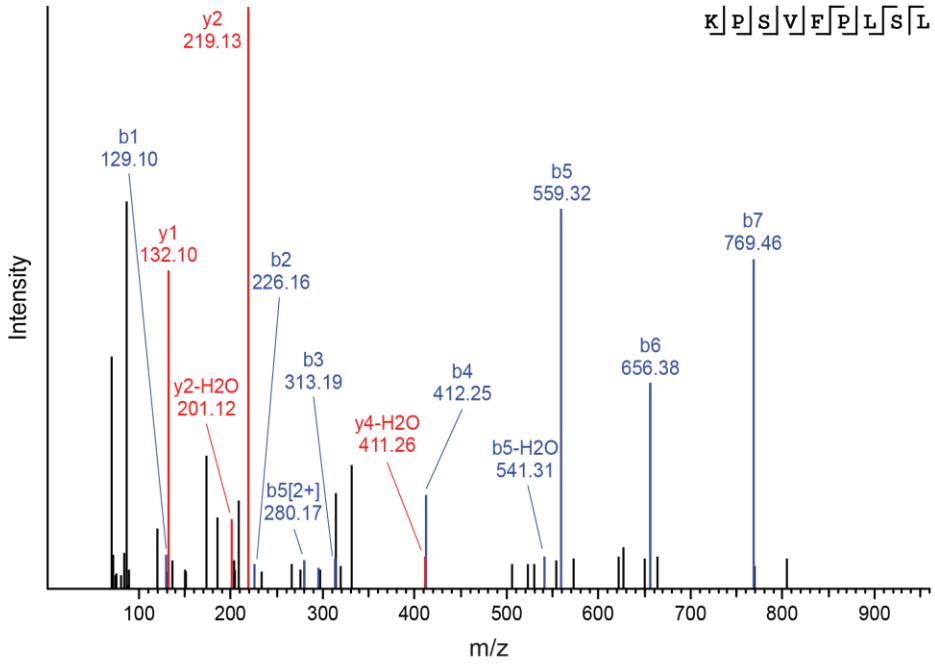
Synthetic peptide



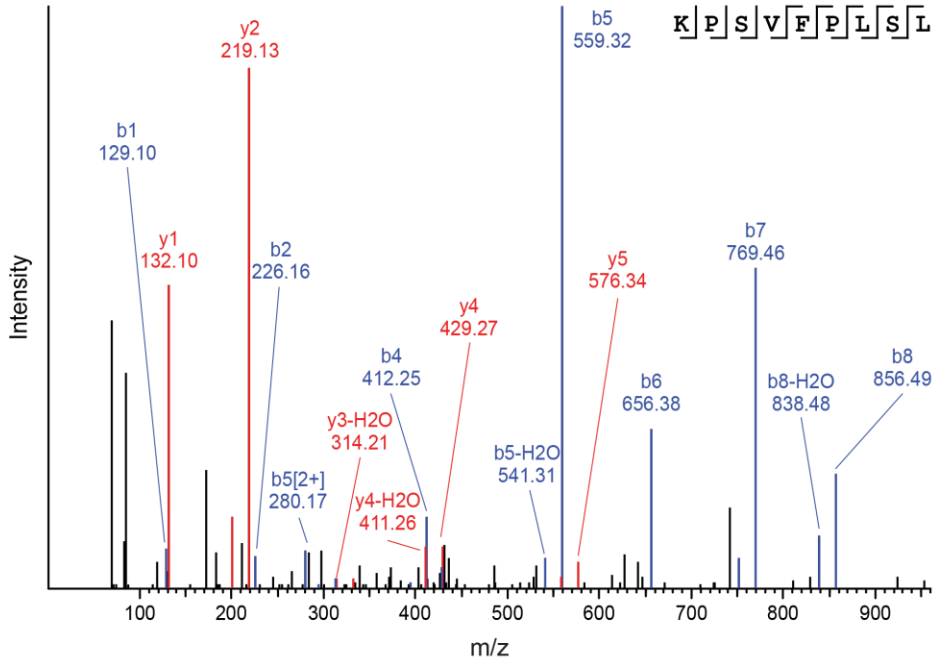
M

KPSVFPLSL - aeTSA

Endogenous peptide



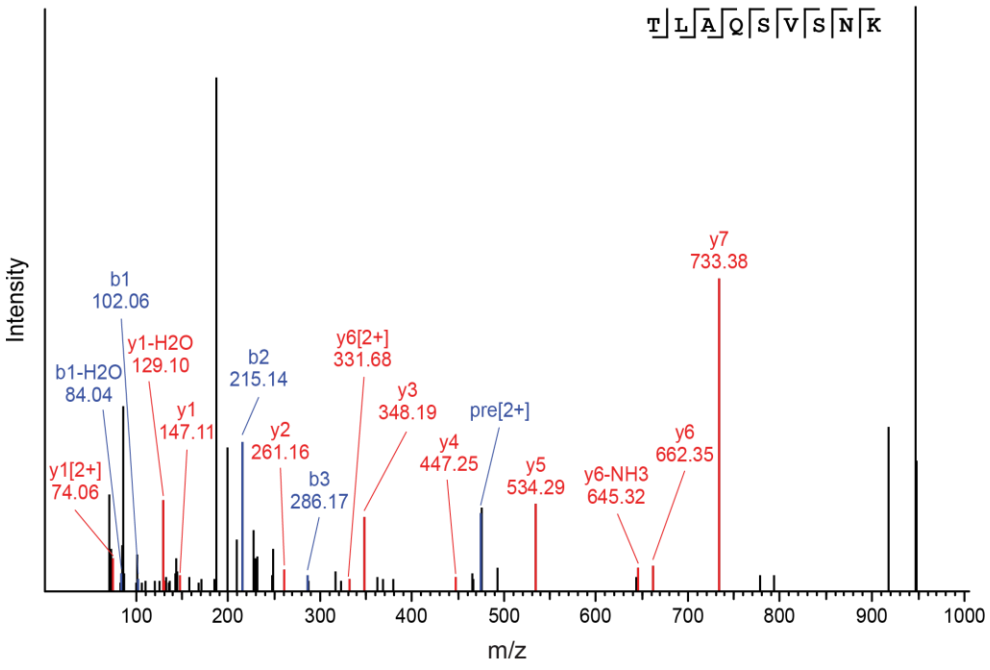
Synthetic peptide



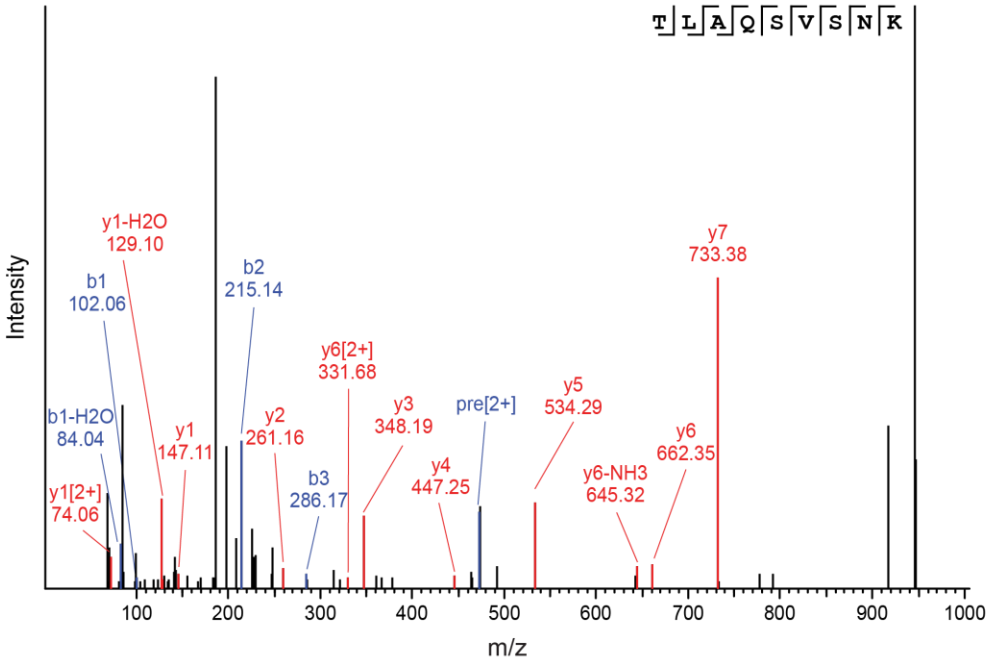
N

TLAQSVSNK - aeTSA

Endogenous peptide



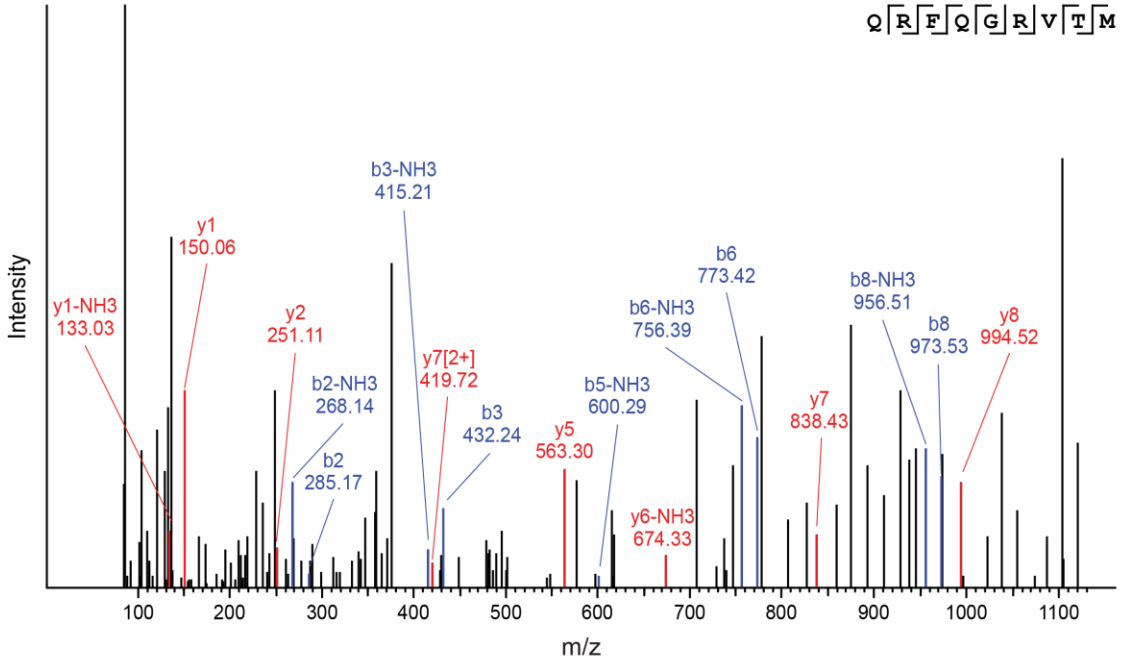
Synthetic peptide



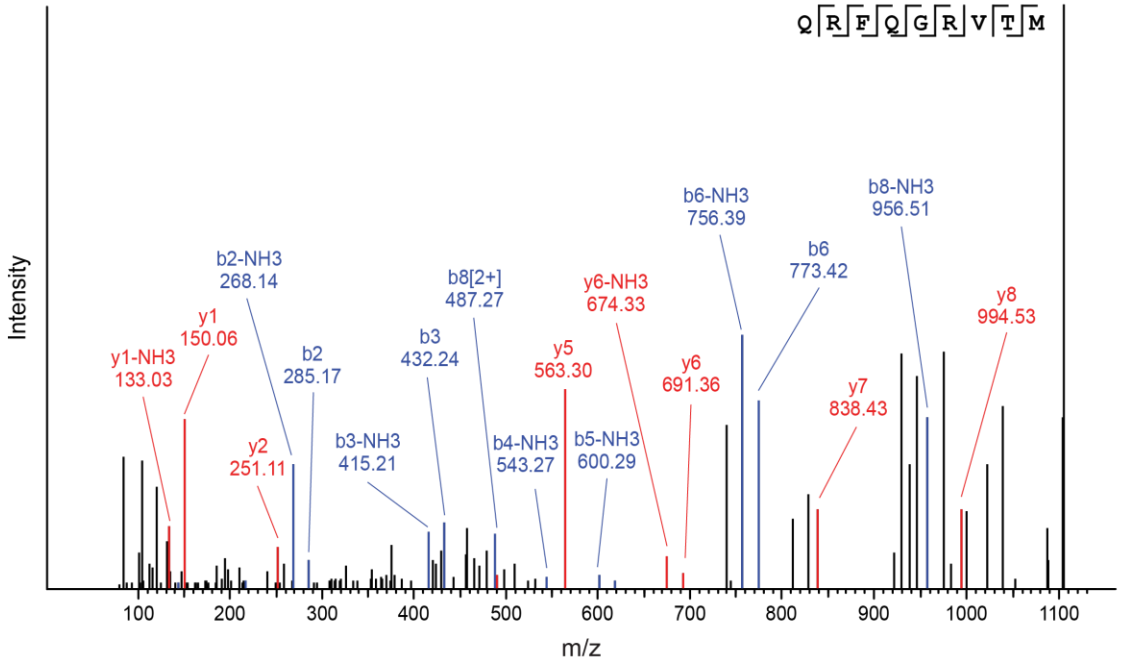
O

QRFQGRVTM - mTSA

Endogenous peptide



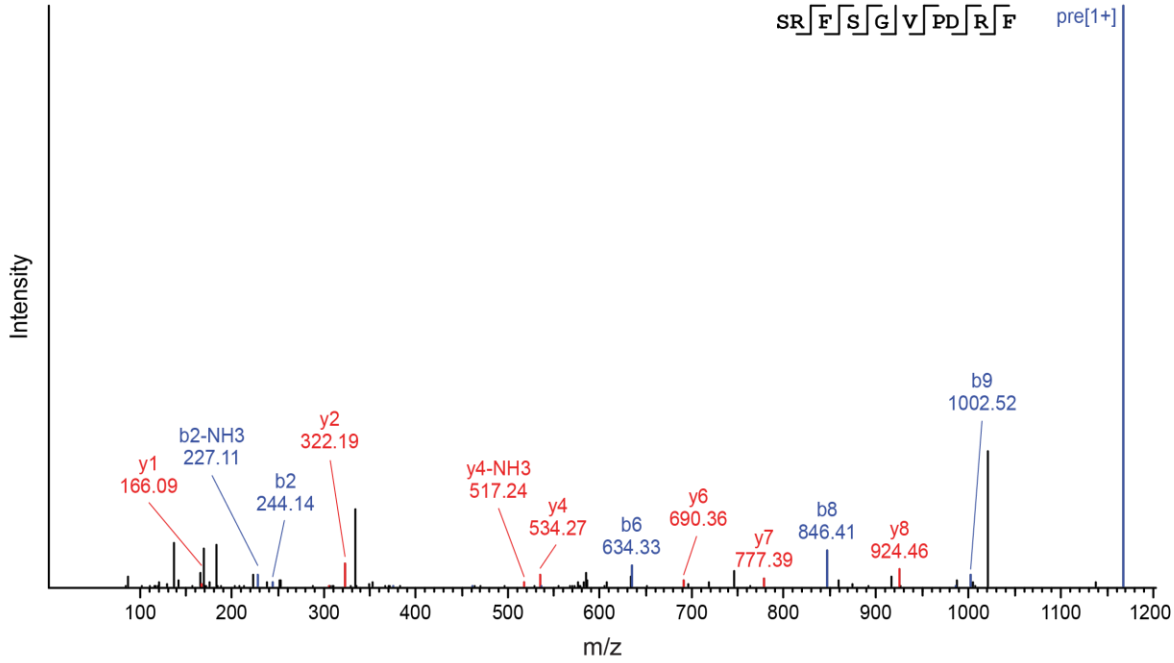
Synthetic peptide



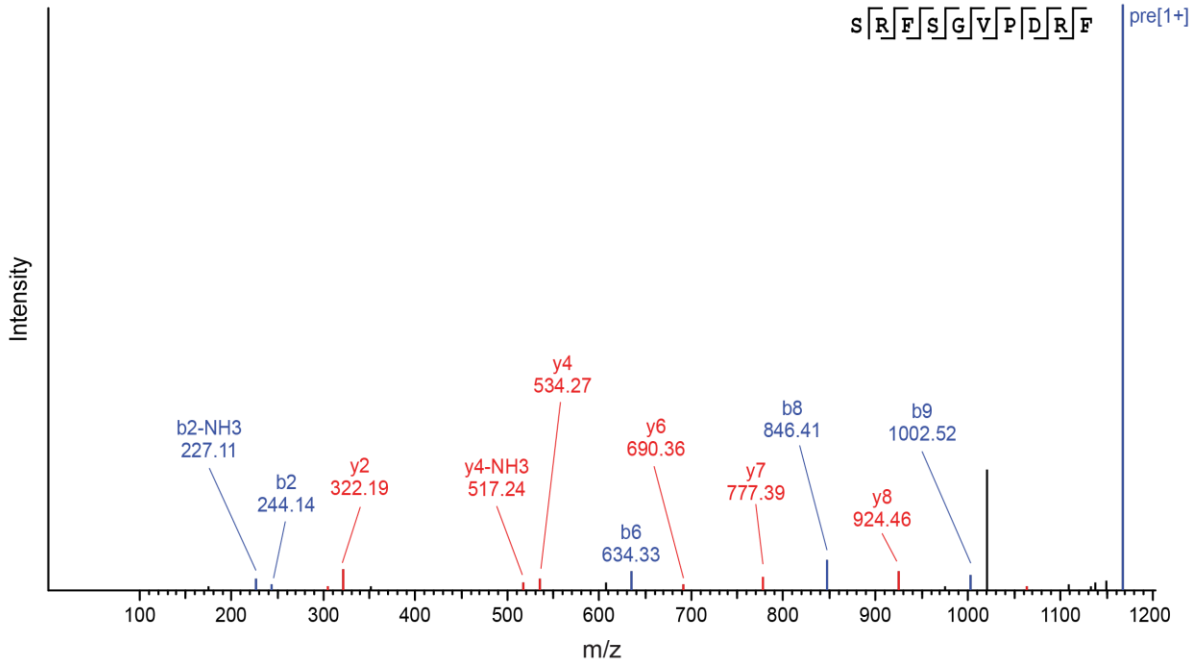
P

SRFSGVPDRF - aeTSA

Endogenous peptide



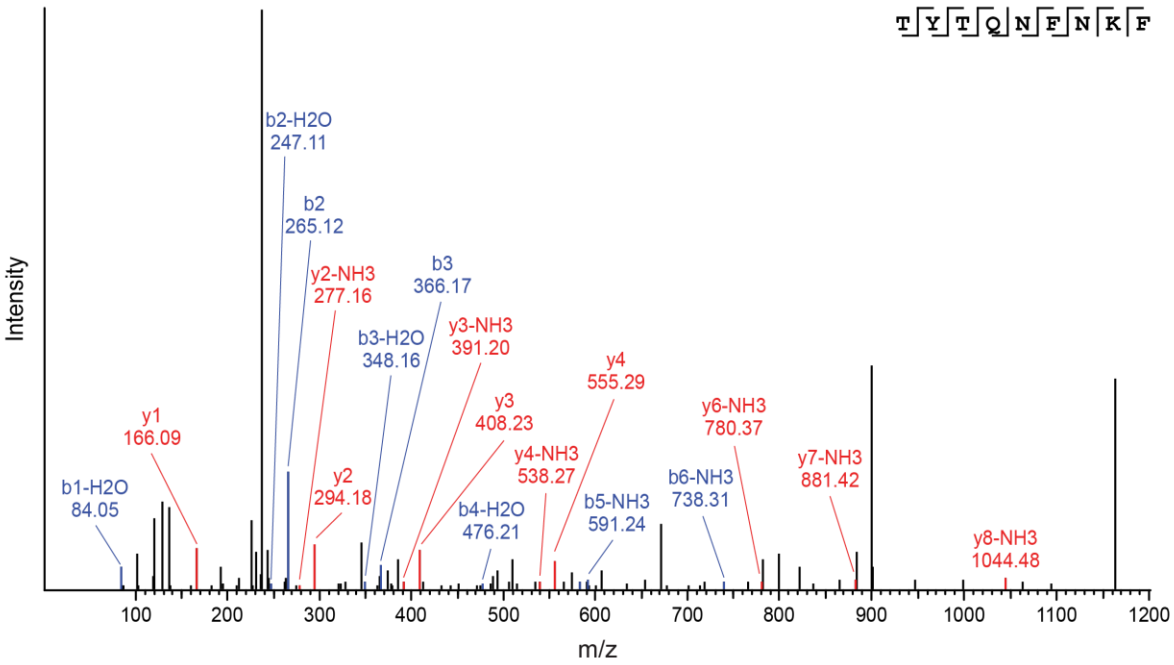
Synthetic peptide



Q

TYTQNFNKF- mTSA

Endogenous peptide



Synthetic peptide

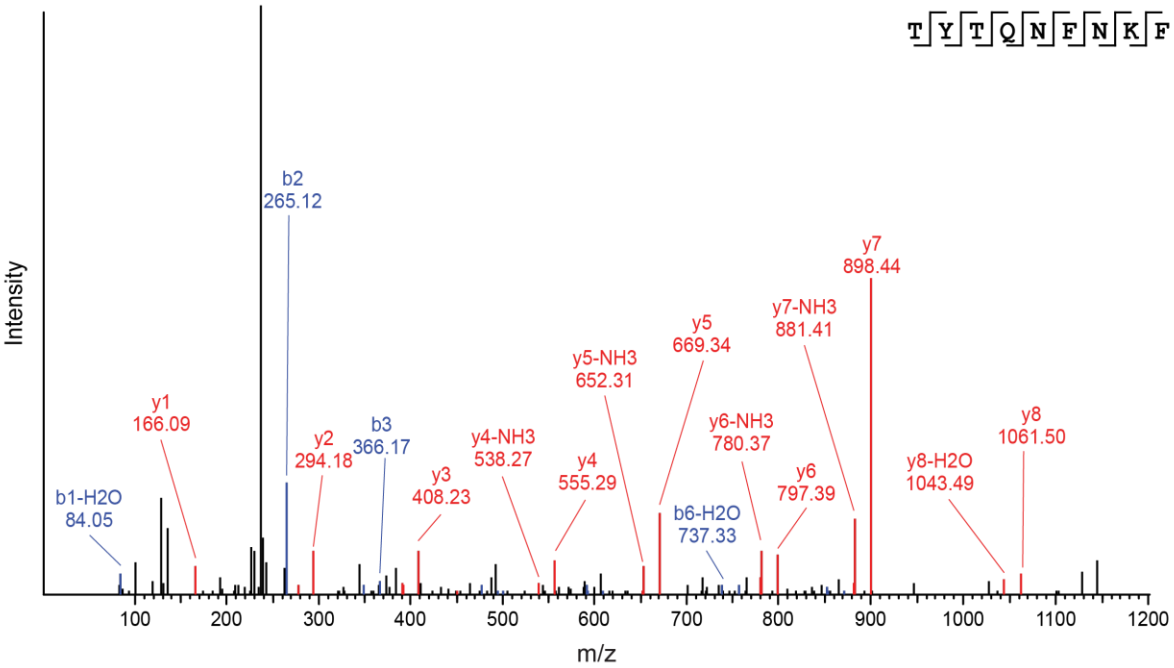


Fig. S11. MS validation of lung cancer TSA candidates using synthetic analogs. Synthetic and endogenous MS/MS spectra for TSA candidates identified in each of our three lung cancers: (**A-L**) lc2, (**M-N**) lc4 and (**O-Q**) lc6. See section **MS validation of TSA candidates** of the **Supplementary Materials** for details.

See associated Excel file for **tables S1-S18**.