Proteoform identification by combining RNA-Seq and top-down mass spectrometry (Supplementary Material)

Wenrong Chen[1] and Xiaowen Liu[1,2]*

[1] Department of BioHealth Informatics, Indiana University-Purdue University Indianapolis, Indianapolis, IN 46202, USA, [2] Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, IN 46202, USA

*Corresponding author: 535 W. Michigan St. Indianapolis IN 46202, xwliu@iupui.edu

1. Parameters and commands for software tools used in RNA-seq data analysis S-2

2. Figures S-4

Figure S1. The five types of alternative splicing events S-4

Figure S2. The distributions of the 8 types of variants reported by ANNOVAR S-5

Figure S3. The distributions of the numbers of nonsynonymous substitutions and indels per transcript reported by ANNOVAR S-6

Figure S4. An example PrSM of the protein ISG15 (ENSEMBL ID: ENSP00000368699) identified by the SWISS-M database. S-7

Figure S5. An example PrSM of the protein EDF1 (ENSEMBL ID: ENSP00000224073) identified by the SWISS-M database. S-8

Figure S6. The two PrSMs containing two heterozygous variants for each sequence. S-9

Figure S7. An example PrSM of the protein KDELR2 (ENSEMBL ID: ENSP00000258739) identified by the BASIC-ES database. S-10

3. Table S-11

Table S1: Parameter settings for TopPIC S-11

Table S2: Proteoforms covering SNV sites identified from SWISS-M, BASIC-M and COMP-M.xlsx.

Table S3: Proteoforms covering SNV sites idenified from BASIC-ES.xlsx.

Table S4: Peptides covering SNV sites identified in bottom-up MS.xlsx.

1. Parameters and commands for software tools used in RNA-seq data analysis

Alignment (STAR 2.7.0c):

STAR --runMode genomeGenerate --genomeDir ./ --genomeFastaFiles Homo_sapiens_assembly38.fasta --runThreadN 30

STAR --genomeDir ./ --readFilesIn /path/to/fastq --runThreadN 30

STAR --runMode genomeGenerate --genomeDir ./ --genomeFastaFiles Homo_sapiens_assembly38.fasta --sjdbFileChrStartEnd SJ.out.tab --sjdbOverhang 101(the read length) --runThreadN 30

Add read groups and sort by coordinates (Picard 2.18.26):

java -jar picard.jar AddOrReplaceReadGroups I=input.sam O=output.bam SO=coordinate RGLB=library RGPL=illumina RGPU=machine RGSM=sample_name

Remove duplicates (Picard 2.18.26):

java -jar picard.jar MarkDuplicates I=../8_read_groups/input.bam O=output.bam CREATE_INDEX=true VALIDATION_STRINGENCY=SILENT M=output.metric

Split'N'Trim (GATK 4.1.0.0):

java -jar gatk-package-4.1.0.0-local.jar SplitNCigarReads -R Homo_sapiens_assembly38.fasta -I input.bam -O split.bam

Base recalibration (GATK 4.1.0.0):

java -jar gatk-package-4.1.0.0-local.jar BaseRecalibrator -R Homo_sapiens_assembly38.fasta -I split.bam --known-sites Homo_sapiens_assembly38.dbsnp138.vcf --known-sites Mills_and_1000G_gold_standard.indels.hg38.vcf --known-sites 1000G_phase1.snps.high_confidence.hg38.vcf -O recalibration_report.grp

(The above known sites are downloaded from public resources.)

java -jar gatk-package-4.1.0.0-local.jar ApplyBQSR -R Homo_sapiens_assembly38.fasta -I split.bam --bqsr-recal-file recalibration_report.grp -O bqsr_split.bam

Variant calling (GATK 4.1.0.0):

java -jar gatk-package-4.1.0.0-local.jar HaplotypeCaller -R Homo_sapiens_assembly38.fasta -I input.bam --dont-use-soft-clipped-bases -stand-call-conf 20.0 -O variant_call.vcf

Variant filtering (GATK 4.1.0.0):

java -jar gatk-package-4.1.0.0-local.jar SelectVariants -R Homo_sapiens_assembly38.fasta -V variant_call.vcf --select-type-to-include SNP --select-type-to-include INDEL --exclude-filtered -O variant_selected.vcf

java -jar gatk-package-4.1.0.0-local.jar VariantFiltration -R Homo_sapiens_assembly38.fasta -V variant_selected.vcf -window 35 -cluster 3 -O variant_filter.vcf

Variant annotation (ANNOVAR, April 16, 2018):

Convert vcf file to annovar input file:

```
convert2annovar.pl -format vcf4 variant_filter.vcf > variant_filter.avinput
```

Download required databases:

```
annotate_variation.pl -downdb -build hg38 seq humandb/hg38_seq/

annotate_variation.pl -downdb wgEncodeGencodeBasicV28 humandb/ -build hg38
```

Filter the annotation database:

```
awk '!/KI270/ && !/GL000/ && !/_alt/' hg38_wgEncodeGencodeBasicV28.txt >
hg38_wgEncodeGencodeBasicV28_filtered.txt
```

Retrieve sequences to mRNA:

```
retrieve_seq_from_fasta.pl -format genericGene -seqfile humandb/hg38_seq/hg38.fa -outfile
humandb/hg38_wgEncodeGencodeBasicV28_filteredMrna.fa
humandb/hg38_wgEncodeGencodeBasicV28_filtered.txt
```

Annotate the variants:

```
annotate_variation.pl -build hg38 -out ex1 -dbtype wgEncodeGencodeBasicV28_filtered
variant_filter.avinput ./annovar/humandb/
```

2. Figures

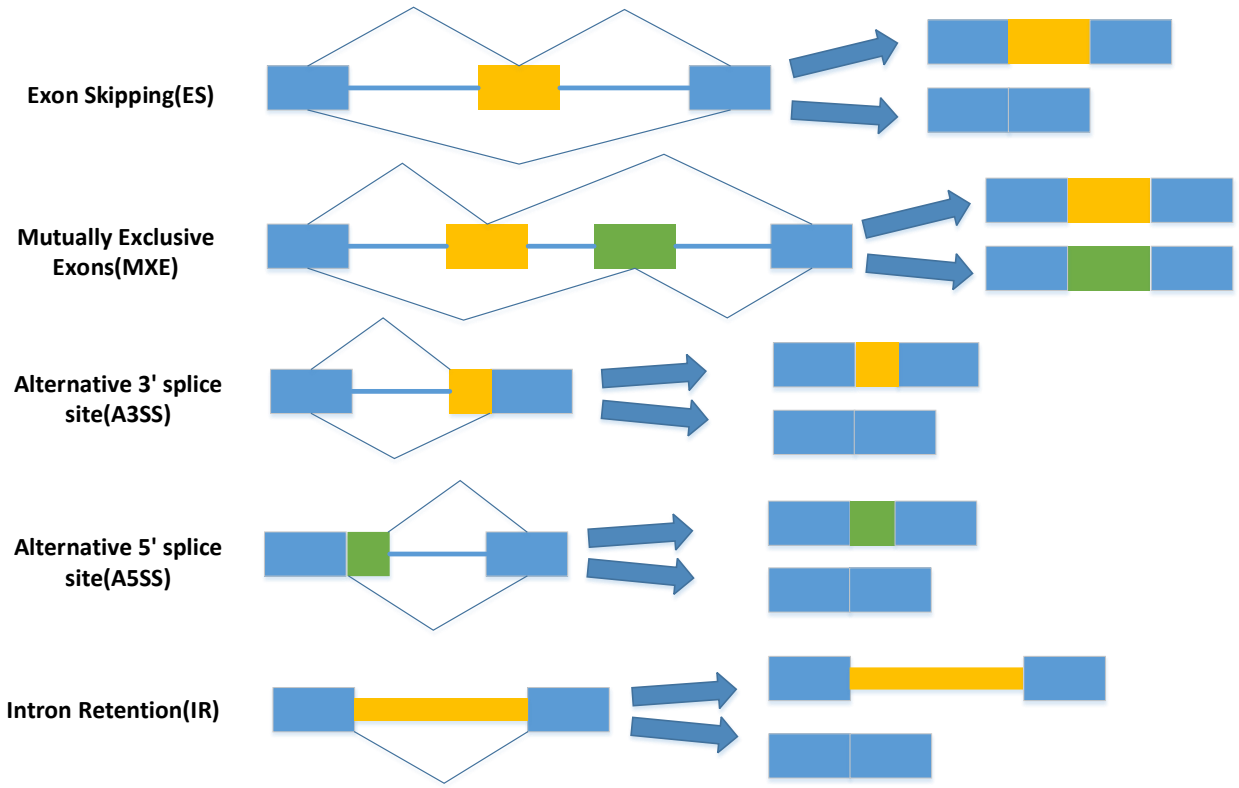

Figure S1. The five types of alternative splicing events: exon skipping, mutually exclusive exons, alternative 3' splice sites, alternative 5' splice sites and intron retention. For each splicing event, it has two transcript isoforms.

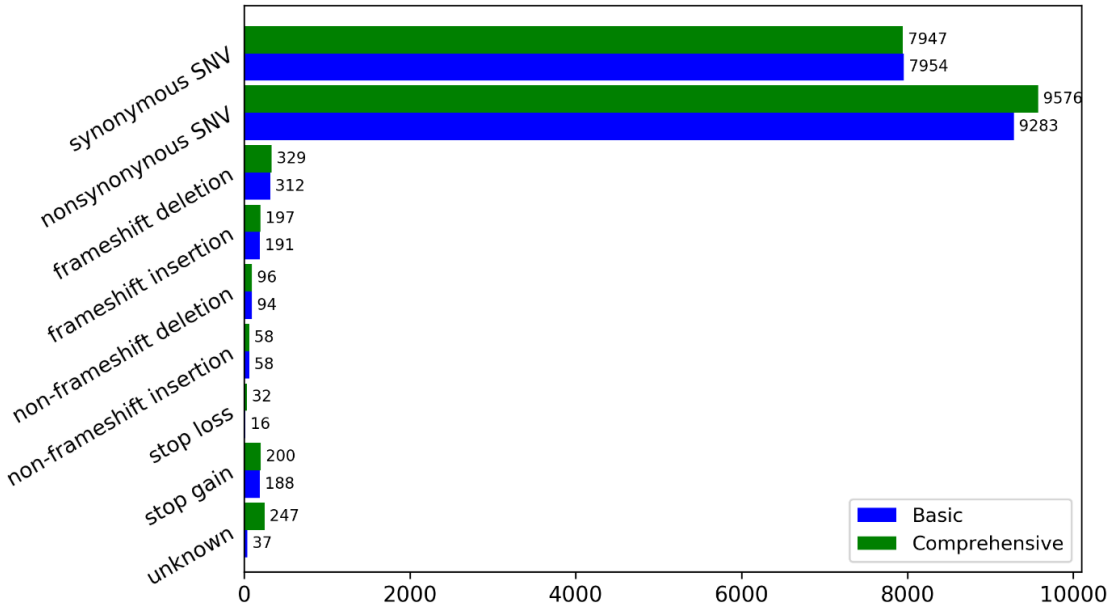

Figure S2. The distributions of the 8 types of variants reported by ANNOVAR on the transcript level with basic and comprehensive annotations. There are 18133 and 18682 variants reported from the basic and comprehensive annotations, respectively.

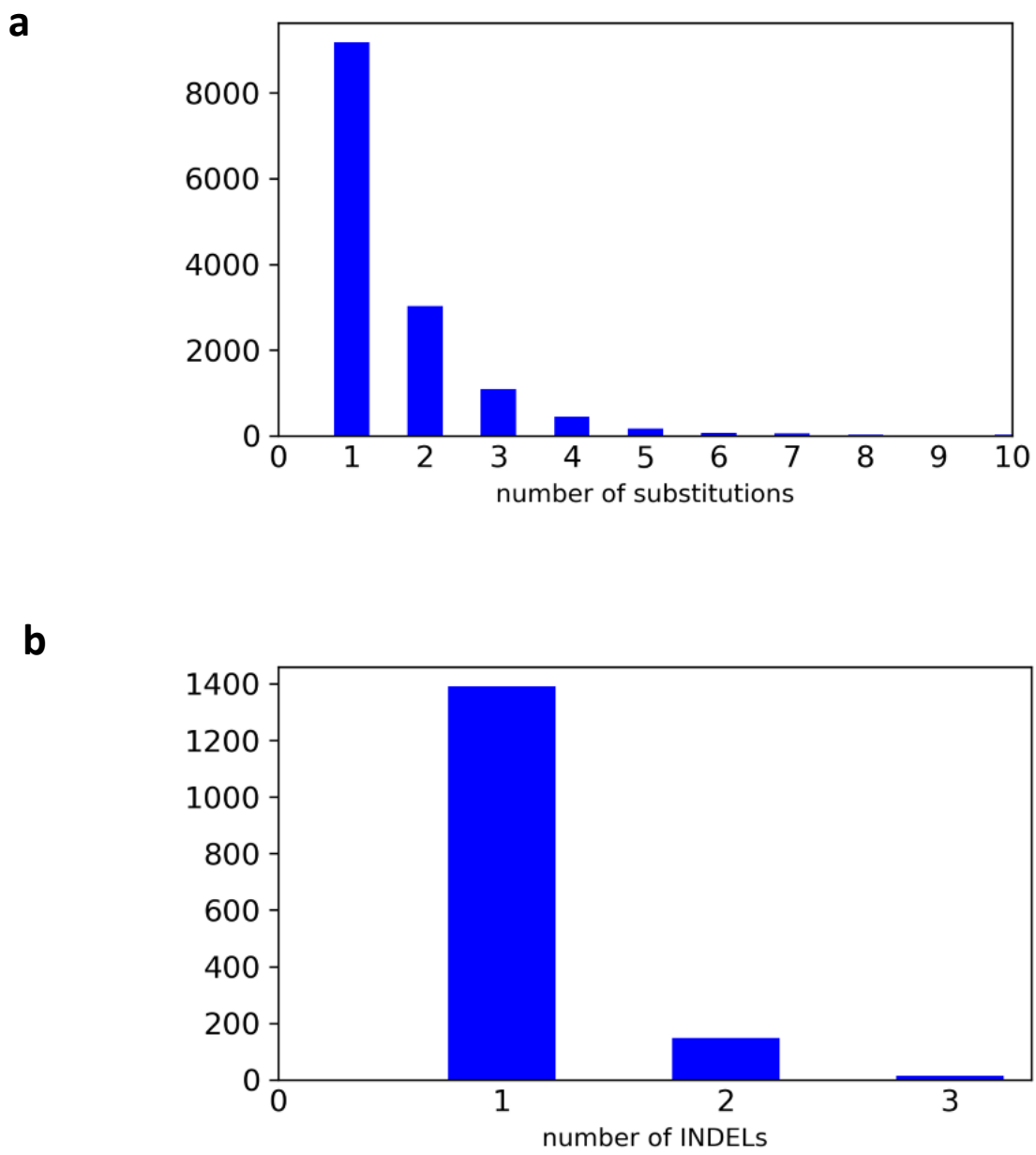

Figure S3. The distributions of the numbers of nonsynonymous substitutions and indels per transcript reported by ANNOVAR from the DLD-1 data set with the GENCODE basic annotation of hg38. (a) The distribution of the number of nonsynonymous substitutions per transcript. (b) The distribution of the number of indels per transcript.

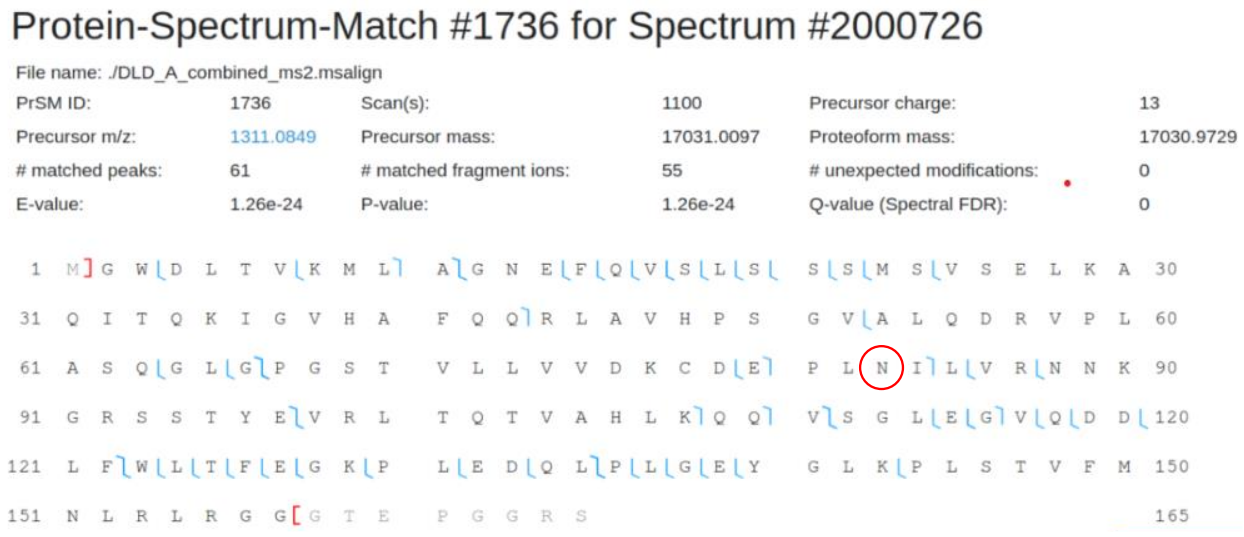

Figure S4. An example PrSM of the protein ISG15 (ENSEMBL ID: ENSP00000368699) identified by the SWISS-M database. The proteoform contains a confidently identified SNV site at position 83 (S -> N).

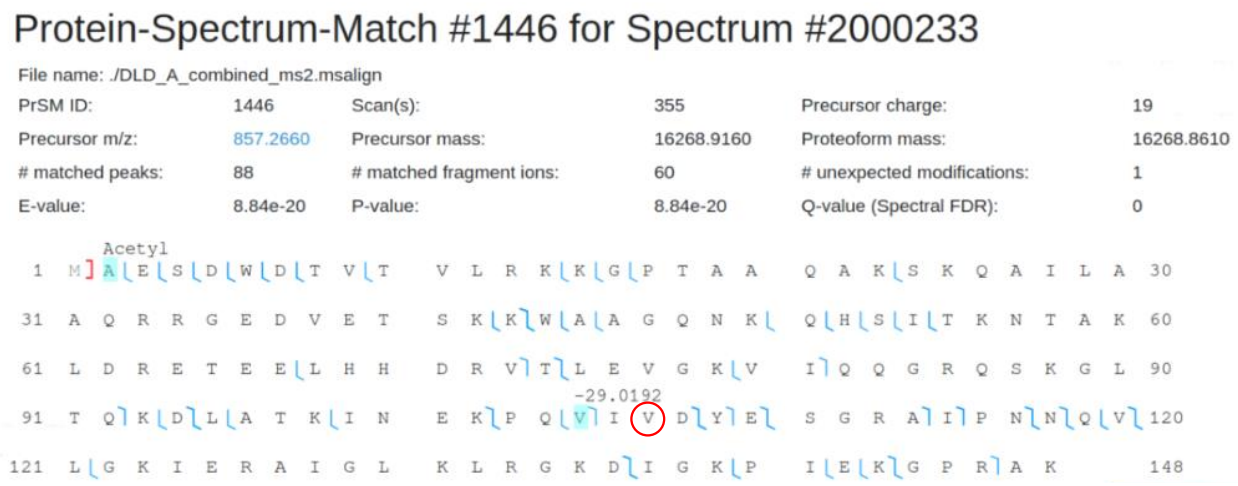

Figure S5. An example PrSM of the protein EDF1 (ENSEMBL ID: ENSP00000224073) identified by the SWISS-M database. The proteoform contains a SNV site at position 107 (A-> V), which is not confidently identified because of the existence of a mass shift -29.0192 near the site.

a

Protein-Spectrum-Match #39 for Spectrum #194

File name: ./DLD_A_combined_ms2.msalign

PrSM ID:	39	Scan(s):	298	Precursor charge:	9
Precursor m/z:	930.0537	Precursor mass:	8361.4176	Proteoform mass:	8361.6275
# matched peaks:	37	# matched fragment ions:	28	# unexpected modifications:	1
E-value:	8.53e-14	P-value:	8.53e-14	Q-value (Spectral FDR):	0

```
... 180 amino acid residues are skipped at the N-terminus ...

181  L P K Q E M A S A S   S S Q R G R R G S G   N F G G G R G D G F  210
                                                            114.1293
211  G G N D N F G R G G   N F S G R G G F G G   S C G G G G Y] G G S  240

241  G D G Y⌊N⌊G⌊F⌊G⌊N⌊D⌊  G⌊S⌊N⌊F⌊G⌊G⌊G⌊G⌊S⌊Y⌊  N⌊D⌊F⌊G⌊N⌊Y⌊N⌊N⌊Q⌊S⌊ 270

271  S N F G P M K G G N   F G G R S S G P Y G   G G G Q Y F A K P R  300

301  N Q G G Y G G S S S⌉  S S S Y G S G R R F                         320
```

b

Protein-Spectrum-Match #1527 for Spectrum #2000471

File name: ./DLD_A_combined_ms2.msalign

PrSM ID:	1527	Scan(s):	715	Precursor charge:	9
Precursor m/z:	907.4746	Precursor mass:	8158.2056	Proteoform mass:	8158.1306
# matched peaks:	16	# matched fragment ions:	14	# unexpected modifications:	1
E-value:	0.26	P-value:	0.23	Q-value (Spectral FDR):	1.86e-03

```
1   M] T⌊P G K H S G A S   A R A A N⌊A G⌊A⌊W⌊G   Y R D F R G G Q K K  30

31  G W W T T P Q L V A⌉  T M P V S⌊P A G⌊S⌊H   K Q Q N F G L N N A⌊ 60
                                   -14.9939
61  T Q⌉P K K S I S F F   A T⌉M K A T⌉S⌉V[K G   Y T G A N Q S R M A  90

91  V S K T V L I P P E   L K T V E K P N P N   I K T T Q V F D I N 120

... 425 amino acid residues are skipped at the C-terminus ...
```

Figure S6. The two PrSMs containing two heterozygous variants for each sequence. (a) The PrSM of the protein HNRNPA1L2 (ENSEMBL ID: ENSP00000350090) identified by BASIC-C. The proteoform contains two confidently identified SNV sites at position 300(Q->R) and 307(V->G). (b) The PrSM of the protein WDR78 (ENSEMBL ID: ENSP00000360062) identified by the BASIC-C. The proteoform contains two confidently identified SNV sites at position 16(G->A) and 33(C->W).

Protein-Spectrum-Match #1900 for Spectrum #2001524

File name: ./DLD_A_combined_ms2.msalign

PrSM ID:	1900	Scan(s):	2297	Precursor charge:	22
Precursor m/z:	1110.3723	Precursor mass:	24406.0314	Proteoform mass:	24406.0392
# matched peaks:	139	# matched fragment ions:	87	# unexpected modifications:	0
E-value:	9.95e-33	P-value:	9.95e-33	Q-value (Spectral FDR):	0

```
  1 M N I F R L T G D L   S H L A A I V I L L   L K I W K T R S C A  30
 31 G I S G K S Q L L F   A L V F T T R Y L D   L F T S F I S L Y N  60
 61 T S M K V I Y L A C   S Y A T V Y L I Y L   K F K A T Y D G N H  90
 91 D T F R V E F L V V   P V G G L S F L V N   H D F S P L E I L W 120
121 T F S I Y L E S V A   I L P Q L F M I S K   T G E A E T I T T H 150
151 Y L F F L G L Y R A   L Y L V N W I W R F   Y F E G F F D L I A 180
181 V V A G V V Q T I L   Y C D F F Y L Y I T   K V L K G K K L S L 210
211 P A                                                             212
```

Figure S7. An example PrSM of the protein KDELR2 (ENSEMBL ID: ENSP00000258739) identified by the BASIC-ES database. The proteoform covers two splicing events reported by MATS. The blue box is the range of the skipped exon ENSE00000923872, and the orange box is the range of the skipped exon ENSE00003531130.

3. Table

Table S1: Parameter settings for TopPIC

Parameter	Value
Number of combined spectra	1
Fragmentation method	FILE
Search type	TARGET+DECOY
Fixed modifications	None
Use TopFD feature file	TRUE
Maximum number of unexpected modifications	1
Error tolerance	15 ppm
Spectrum-level cutoff type	FDR
Spectrum-level cutoff value	0.01
Proteoform-level cutoff type	FDR
Proteoform-level cutoff value	0.01
Allowed N-terminal forms	NONE, NME, NME_ACETYLATION, M_ACETYLATION
Maximum mass shift of modifications	500 Da
Minimum mass shift of modifications	-500 Da
Thread number	1
E-value computation	Generation function