De Novo sequencing-assisted homology search for DIA data analysis enables low abundance peptide variants discovery

Rui Qiao[1†], Hechen Li[1†], Haibo Bian[1], Lei Xin[1], Baozhen Shan[1*]

[1]Bioinformatics Solutions Incorporations, Waterloo, Canada
[†]These authors contributed equally
*Corresponding author. E-mail: bshan@bioinfor.com

Abstract

Data-independent acquisition mass spectrometry (DIA-MS) has emerged as a powerful approach for comprehensive proteome profiling. Spectral library search and library-free search are the two major approaches for DIA data analysis. The spectral library search requires high-quality spectral libraries derived from the search results of data-dependent acquisition (DDA) experiments, while library-free approaches rely on prediction models to generate in silico libraries. Both methodologies constrain the search space to the peptide list in the database, limiting the discovery of variant peptides arising from genetic variations or mutations. We present a novel computational method DIAVariant designed to identify peptide sequence variants directly and solely from complex DIA spectra while rigorously controlling the false discovery rate. Our experimental results demonstrate that DIAVariant successfully identifies sequence variants previously detected through proteogenomic approaches, while maintaining high specificity across multiple datasets. When integrated with existing DIA database search solutions, our approach constitutes a comprehensive analytical workflow capable of identifying peptides both represented within reference protein databases and those arising from sequence variations not captured in standard databases.

1 Introduction

Data-independent acquisition mass spectrometry (DIA-MS) has revolutionized the field of proteomics, rapidly gaining prominence as a powerful and comprehensive technique for protein identification and quantification. Unlike traditional data-dependent acquisition (DDA) methods, which selectively fragment and analyse the most abundant precursor ions in a sample, DIA-MS systematically fragments all ions within a defined mass range during each scan cycle. This comprehensive fragmentation approach offers several advantages over DDA, leading to its increasing adoption in the proteomics community. Specifically, DIA-MS

provides improved quantitative accuracy, reproducibility, and higher proteome coverage [1, 2].

The increasing popularity of DIA-MS can be attributed to several key factors. Firstly, its ability to generate more complete and reproducible quantitative data allows for more reliable comparison of protein expression levels across different samples or conditions. Secondly, the inherent nature of DIA, where all fragment ions are recorded, allows for deeper proteome coverage than DDA. This is because low-abundance peptides, which might be missed by the selective nature of DDA, are more likely to be identified and quantified in a DIA experiment. Finally, the development of sophisticated data analysis tools and workflows has made the processing and interpretation of DIA data more accessible [3, 4, 5], further driving its widespread adoption.

Sequence variants in proteins, arising from single nucleotide polymorphisms (SNPs), insertions, deletions, and other genetic mutations, play a crucial role in various biological processes and disease. These variants can alter protein structure, function, stability, and interactions, potentially leading to phenotypic differences, drug resistance, or disease susceptibility [6, 7, 8, 9]. Identifying sequence variants at the protein level is therefore critical for understanding the molecular basis of human disease, as well as for developing personalized medicine strategies.

Despite the inherent advantages of DIA-MS, the identification of sequence variants from DIA data presents substantial methodological and computational challenges. Current DIA analysis methodologies predominantly rely on spectral matching against libraries derived from DDA experiments or in silico predictions based on genomic databases [3, 4]. This library-dependent paradigm inherently constrains the discovery of novel peptides absent from the reference libraries, particularly those harbouring sequence variants. While proteogenomic approaches that integrate genomic and proteomic data can facilitate the construction of comprehensive variant-aware libraries [10], these methodologies typically require specialized data acquisition protocols and sequential data analysis pipelines which might carry mistakes from upstream genomics data over to downstream proteomics data analysis. Moreover, recent investigations in neoantigens have revealed that sequence variants may originate from alternative splicing events and previously unannotated non-coding regions [11, 12], which lie beyond the scope of conventional proteogenomic approaches. Recently, various de novo sequencing methodologies have been proposed for novel peptide identification from DIA data [13, 14]. These approaches frequently employ complex neural network architectures that function as black box models and the predicted sequences lack interpretability. Furthermore, their validation has been predominantly limited to sequences present in existing libraries, leaving considerable uncertainty regarding the false discovery rate among the reported novel peptides. In this work, we present a novel computational framework DIAVariant for the direct identification of sequence variants from DIA-MS data without reliance on prior genomic information. Our approach extends conventional library-dependent methodologies by implementing a two-stage analysis pipeline. First, we leverage the PEAKS DIA algorithm [15], which facilitates the identification of both library-annotated and novel peptides from complex DIA spectra. Subsequently, we employ the SPIDER search algorithm [16] to correct de novo sequencing errors through reference genome alignment, thereby identifying peptide sequences that diverge from reference peptides by only one or two amino acid substitutions. To enhance the specificity of our variant identification, we implement a secondary validation process wherein we predict

the indexed Retention Time (iRT) values and theoretical fragment ion spectra for putative variant sequences. These predictions enable a second round of library searching, effectively filtering out novel sequence precursors that exhibit excessive fragment ion signal overlap with confidently identified reference peptides. This two-stage search strategy will give us a list of novel peptide variants supported by evidence from both liquid chromatography (LC) and MS dimensions, accompanied by statistically meaningful scores that facilitate stringent false discovery rate control. Our experimental validation demonstrates that the proposed methodology successfully identifies sequence variants previously detectable only through proteogenomic approaches.

2 Methods

Our workflow is peptide-centric and begins with MS1 feature detection to identify reliable MS1 signals with well-defined isotopic patterns characteristic of ionized peptides. Subsequently, we perform PEAKS DIA database search against the user-provided reference protein database. Precursors identified at 1% FDR are then mapped to previously detected MS1 features based on retention time and precursor m/z correspondence.

For MS1 features lacking confident peptide identification from the reference database, we implement DIA de novo sequencing [13] to predict peptide sequences that optimally match the MS2 scans associated with each original MS1 feature. The SPIDER algorithm [16] is then applied to confident de novo candidates, utilizing the de novo peptide sequence, MS1 feature information, and reference database to correct potential de novo sequencing errors and identify sequences that could align to the reference protein database.

All SPIDER candidate peptides undergo an initial filtering, retaining only peptides with one or two mutations. We then predict the indexed retention time (iRT) for these potential mutated peptides and convert it to experimental retention time using the RT regression model established during the previous database search. Peptides exhibiting substantial discrepancies between predicted and observed retention times are eliminated from consideration.

To control the quality of the remaining sequence variants, we run another round of spectral library search. This library is constructed on-the-fly by combining the remaining candidate peptide variants with peptides identified in the preceding DIA database search. Retention times in the library are assigned according to the empirical values from previous search results, and a PEAKS DIA spectral library search specific [15] to the data is conducted by directly targeting these retention times. Q-values are reported for each precursor, allowing users to apply appropriate thresholds and manually evaluate the biological significance of reported sequence variants.

For any given MS sample, our workflow reports two sets of identification results. One is the set of sequences from the protein database, with an FDR estimate using the traditional target-decoy search strategy. The other one is a set of sequence variants that might exist in the original biological sample. A flow chart of DIAVariant is demonstrated in Figure 1.

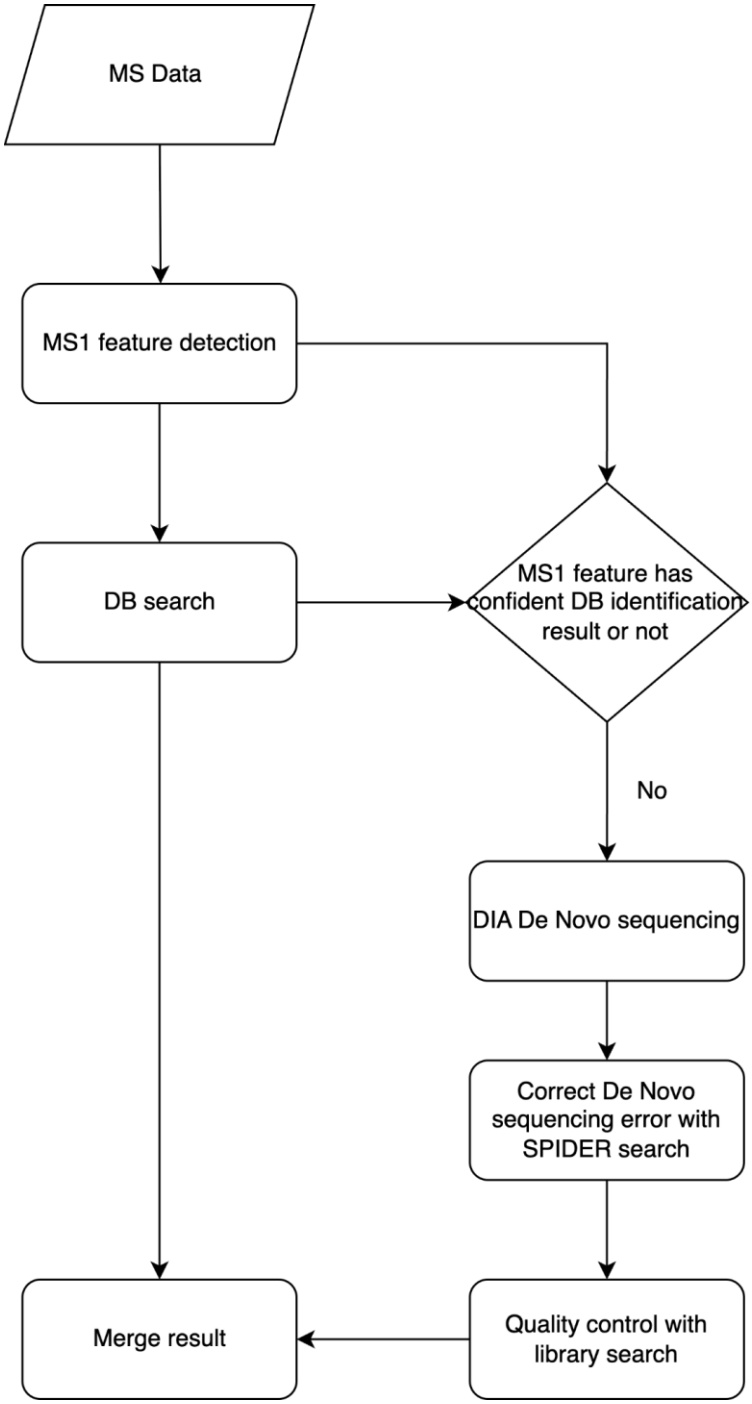


Figure 1. DIAVariant workflow

3 Results

As illustrated in Figure 1, DIA DB search is integrated within our proposed analytical workflow. The implementation of accurate and sensitive DB search methodologies is essential for minimizing false positive identifications in sequence variants. To validate the performance of our DIA DB search engine, we benchmarked PEAKS DIA DB against two recently published datasets [17][18]. Our results in Table 1 demonstrates that the PEAKS

DIA DB search algorithm achieves identification and quantification performance metrics that are at least comparable to those obtained with existing DIA solutions.

Dataset	Software	Number of quantified protein groups	Number of Protein groups with CV < 20%	Number of Protein groups with CV < 10%
PXD046453	PEAKS 13	8655	8401	7549
	Third party software	8828	8436	7589
PXD050030	PEAKS 13	7586	7352	6398
	Third party software	7560	7071	5912

Table 1. PEAKS DIA DB benchmarks

To assess the specificity of sequence variants identified by our workflow, we conducted validation using public MS datasets analysed against non-native reference protein databases from homologous species. For instance, human samples were searched against a mouse protein database. This cross-species validation approach enabled us to evaluate the biological relevance of reported sequence variants by determining the proportion that could be confidently identified (i.e. q-value < 0.01) when searched against the original species' protein database. We conducted the validation on ABRF data [19] (human sample, search against mouse DB) and a recent dataset from Orbitrap Astral [20] (MSV000095360, mouse sample, search against human DB). As demonstrated in Table 2, more than 85% of the peptide variants reported by DIAVariant are confident precursors from the native reference database.

	Number of reported sequence variants	Number of sequence variants that can be confidently identified in the native database	Accuracy
ABRF data	340	315	92.65%
MSV000095360	964	844	87.55%

Table 2. Cross-species validation result

We then applied DIAVariant to analyse the dataset published by Fierro-Monti et al. [10]. In their investigation, Fierro-Monti et al. performed exome sequencing on HeLa cell line samples and conducted variant calling to identify 233 canonical HeLa variant protein sequences. These variant sequences were then incorporated into the human reference protein database to generate a customized protein database. The authors then searched DIA MS spectra from the HeLa samples against this customized database using DIA-NN 1.8.1. Among the sequence variants identified at 1% FDR, Fierro-Monti et al. reported six peptides validated through stable isotope dilution and parallel reaction monitoring. In contrast to the proteogenomic approach employed by Fierro-Monti et al., our proposed

method operates directly and exclusively from MS data, which means confident MS1 precursor isotope envelopes are needed to confirm the presence of peptide variants in the original sample. Upon thorough examination of the raw MS scans, we determined that four of the six previously reported variant peptides lacked sufficient MS1 signals and consequently could not be detected by our method. Notably, for the remaining two peptides with adequate MS1 evidence, our method successfully retrieved them on several DIA MS samples. Table 3 shows a summary of this result. The matching spectra of the two peptides are demonstrated in Supplementary Figure 1-7. In addition to the two previously reported variant peptides, DIAVariant identified 426 additional variants. Several of these variants exhibited low abundance, characterized by low MS1 precursor intensity, and consequently were hard to be detected in DDA data. Supplementary Figure 8 presents a representative example of such low-abundance variants detected by DIAVariant.

Peptide variants reported by proteogenomic method	Our proposed method
LEQDLQQIQAK	Found in 2/5 samples
ERLEQDLQQIQAK	No MS1 signal
QGTFHSQQALEYGTR	No MS1 signal
KIEGDLQGDHQK	No MS1 signal
VYVVPMGGVR	No MS1 signal
NELSGALTGLIR	Found in 5/5 samples

Table 3. Our method finds 2 out of the 6 previous reported sequence variants

4 Discussion

In this study, we present a novel workflow for direct identification of sequence variants exclusively from DIA MS data. Our experimental results demonstrate that DIAVariant successfully identifies previously reported sequence variants detected through proteogenomic approaches, given the presence of robust MS1 signals. In addition, DIAVariants reported hundreds more peptide variants of different abundance. The biological validation of these candidate variants requires further experimental investigation and represents an important avenue for future research.

Notably, our approach exhibits potentials for detecting sequence variants arising from alternative splicing events or non-coding genomic regions—variants that remain largely inaccessible to conventional proteogenomic methodologies. This represents a substantial advancement in variant detection capabilities within proteomics research, addressing critical gaps in existing analytical frameworks.

References

[1] Blackburn, Kevin, et al. "Improving protein and proteome coverage through data-independent multiplexed peptide fragmentation." Journal of proteome research 9.7 (2010): 3621-3637.

[2] Skowronek, Patricia, et al. "Rapid and in-depth coverage of the (phospho-) proteome with deep libraries and optimal window design for dia-PASEF." Molecular & Cellular Proteomics 21.9 (2022).

[3] Demichev, Vadim, et al. "DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput." Nature methods 17.1 (2020): 41-44.

[4] Bernhardt, Oliver M., et al. "Spectronaut: a fast and efficient algorithm for MRM-like processing of data independent acquisition (SWATH-MS) data." Biognosys. ch (2012).

[5] Yu, Fengchao, et al. "Analysis of DIA proteomics data using MSFragger-DIA and FragPipe computational platform." Nature communications 14.1 (2023): 4154.

[6] Hecht, Maximilian, Yana Bromberg, and Burkhard Rost. "Better prediction of functional effects for sequence variants." BMC genomics 16 (2015): 1-12.

[7] Khurana, Ekta, et al. "Role of non-coding sequence variants in cancer." Nature Reviews Genetics 17.2 (2016): 93-108.

[8] Abelson, Jesse F., et al. "Sequence variants in SLITRK1 are associated with Tourette's syndrome." Science 310.5746 (2005): 317-320.

[9] Gudbjartsson, Daniel F., et al. "A sequence variant in ZFHX3 on 16q22 associates with atrial fibrillation and ischemic stroke." Nature genetics 41.8 (2009): 876-878.

[10] Fierro-Monti, Ivo, et al. "Assessment of Data-Independent Acquisition Mass Spectrometry (DIA-MS) for the Identification of Single Amino Acid Variants." Proteomes 12.4 (2024): 33.

[11] Xie, Na, et al. "Neoantigens: promising targets for cancer therapy." Signal transduction and targeted therapy 8.1 (2023): 9.

[12] Park, Jiyeon, and Yeun-Jun Chung. "Identification of neoantigens derived from alternative splicing and RNA modification." Genomics & informatics 17.3 (2019): e23.

[13] Tran, Ngoc Hieu, et al. "Deep learning enables de novo peptide sequencing from data-independent-acquisition mass spectrometry." Nature methods 16.1 (2019): 63-66.

[14] Sanders, Justin, et al. "A transformer model for de novo sequencing of data-independent acquisition mass spectrometry data." bioRxiv (2024): 2024-06.

[15] Xin, Lei, et al. "A streamlined platform for analyzing tera-scale DDA and DIA mass spectrometry data enables highly sensitive immunopeptidomics." Nature Communications 13.1 (2022): 3108.

[16] Han, Yonghua, Bin Ma, and Kaizhong Zhang. "SPIDER: software for protein identification from sequence tags with de novo sequencing error." Proceedings. 2004 IEEE Computational Systems Bioinformatics Conference, 2004. CSB 2004.. IEEE, 2004.

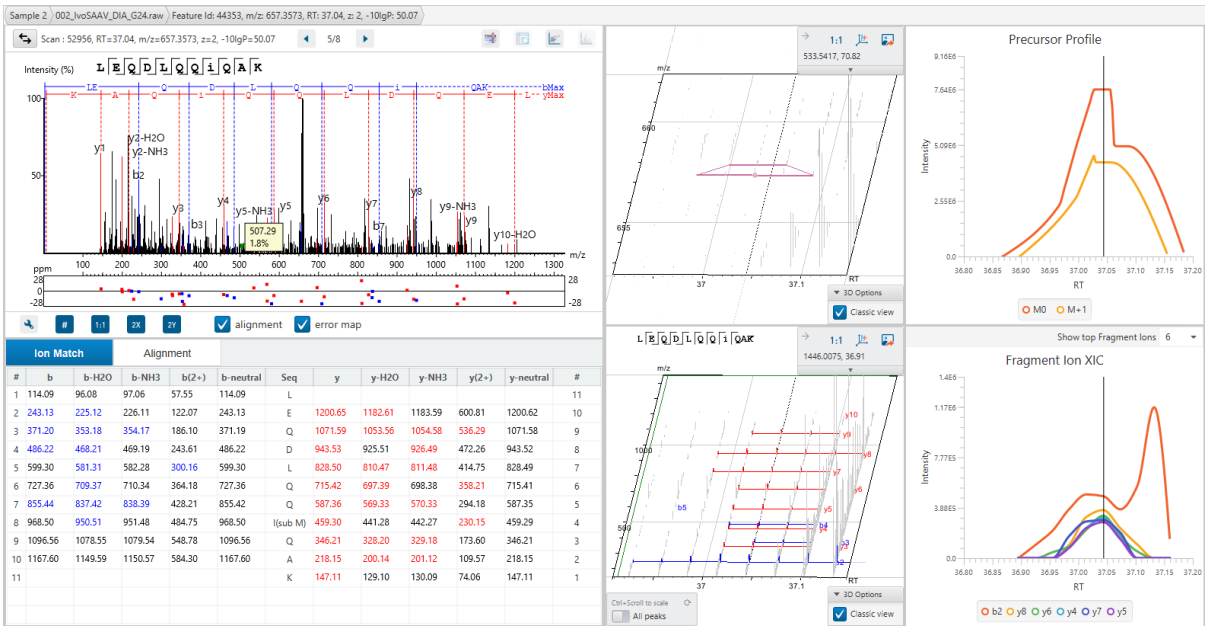
[17] Guzman, Ulises H., et al. "Ultra-fast label-free quantification and comprehensive proteome coverage with narrow-window data-independent acquisition." Nature biotechnology 42.12 (2024): 1855-1866.

[18] Gu, Kongxin, et al. "Improving Proteomic Identification Using Narrow Isolation Windows with Zeno SWATH Data-Independent Acquisition." Journal of Proteome Research 23.8 (2024): 3484-3495.

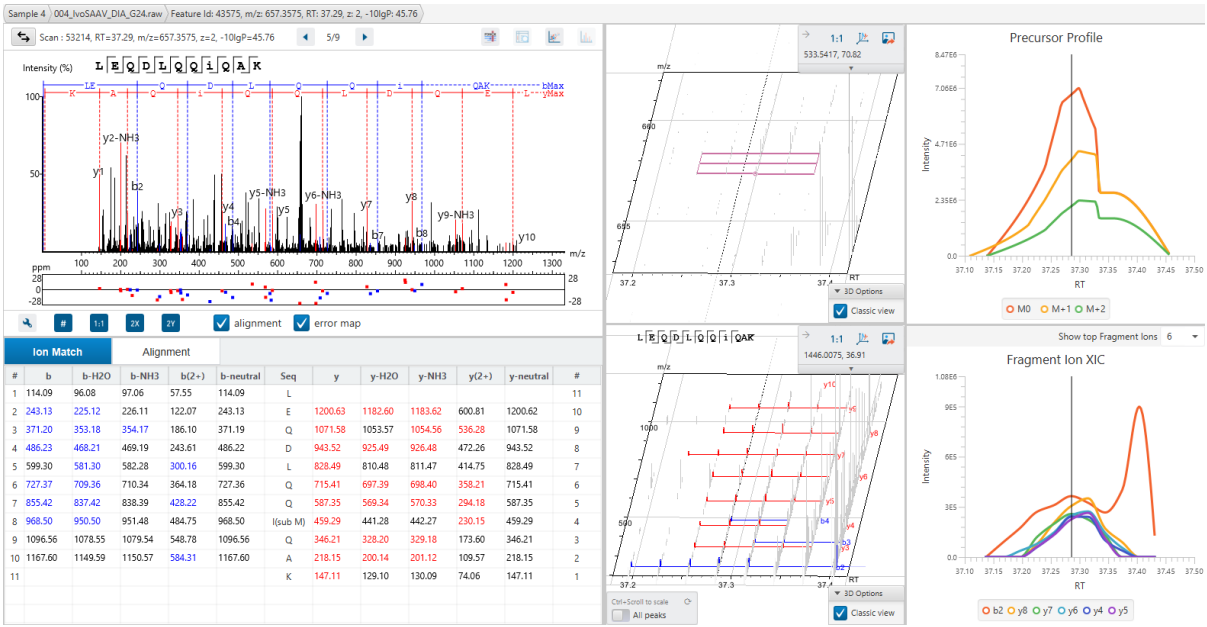
[19] Neely, Benjamin, et al. "Data Independent Acquisition in Practice: Foundations and Resources for Implementing DIA." Journal of Biomolecular Techniques: JBT 30.Suppl (2019): S46.

[20] Hendricks, Nathan G., et al. "An inflection point in high-throughput proteomics with Orbitrap Astral: analysis of biofluids, cells, and tissues." Journal of Proteome Research 23.9 (2024): 4163-4169.

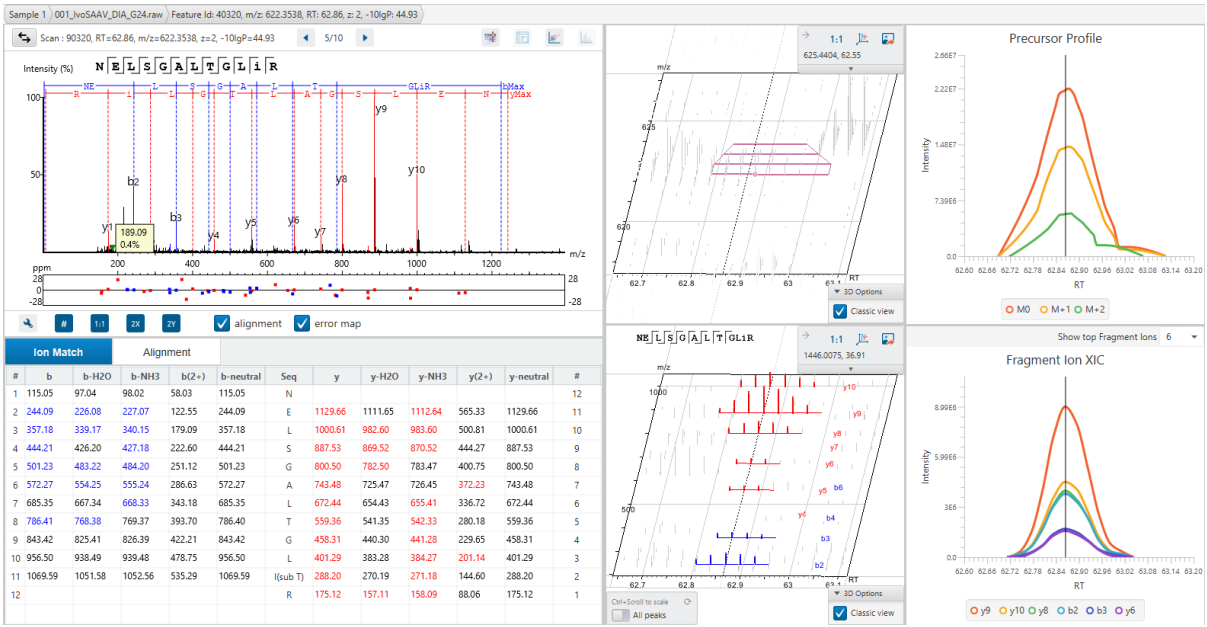
Supplementary information



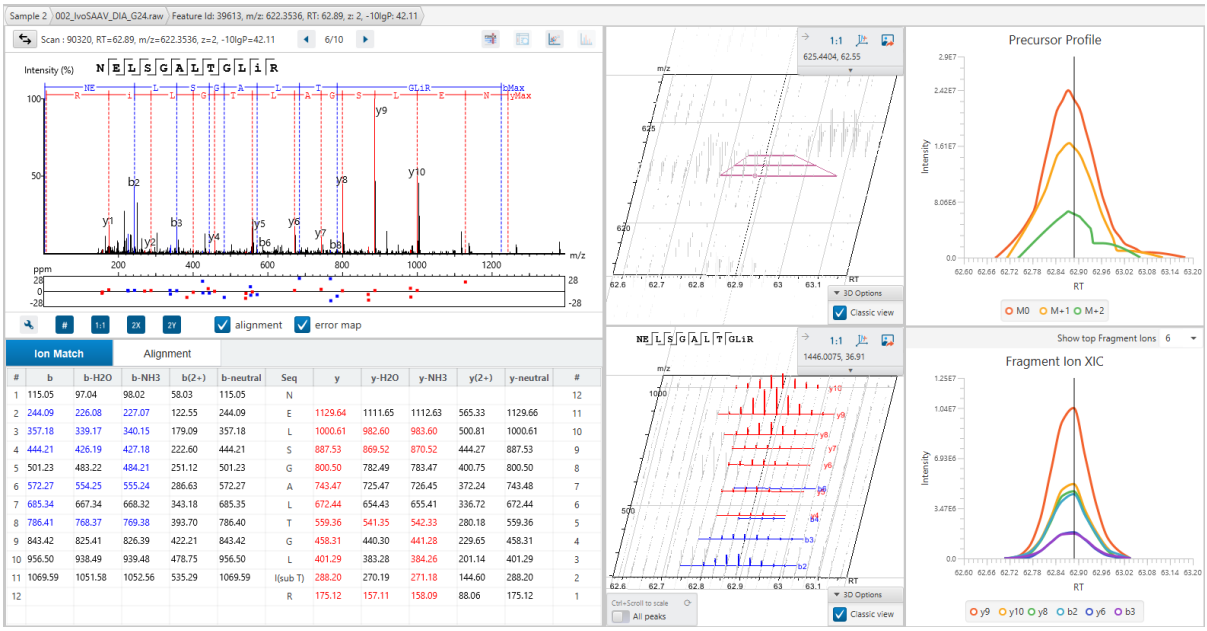
Supplementary Figure 1. LEQDLQQIQAK in Sample 2



Supplementary Figure 2. LEQDLQQIQAK in Sample 4

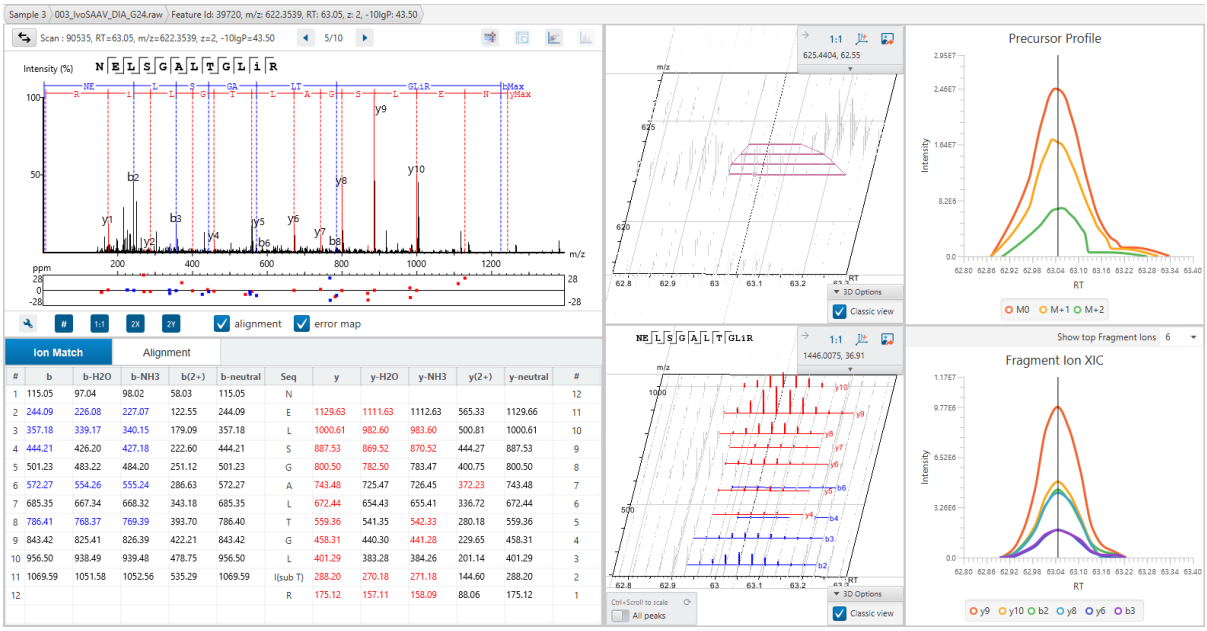


Supplementary Figure 3. NELSGALTGLIR in Sample 1



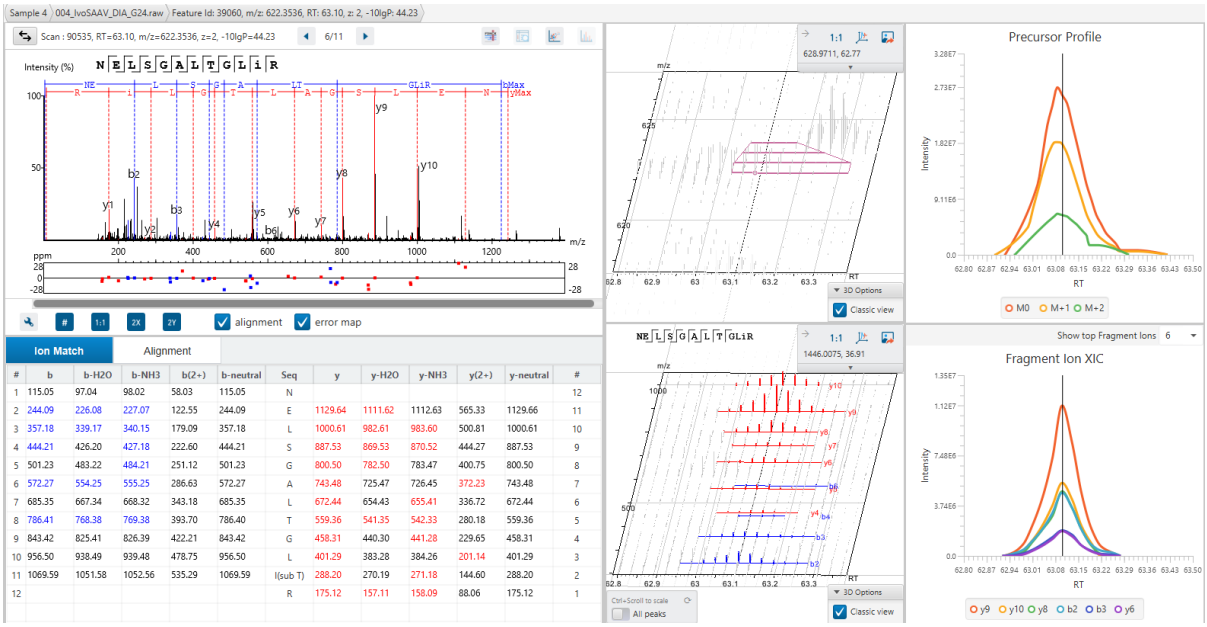
#	b	b-H2O	b-NH3	b(2+)	b-neutral	Seq	y	y-H2O	y-NH3	y(2+)	y-neutral	#
1	115.05	97.04	98.02	58.03	115.05	N						12
2	244.09	226.08	227.07	122.55	244.09	E	1129.64	1111.65	1112.63	565.33	1129.66	11
3	357.18	339.17	340.15	179.09	357.18	L	1000.61	982.60	983.60	500.81	1000.61	10
4	444.21	426.19	427.18	222.60	444.21	S	887.53	869.52	870.52	444.27	887.53	9
5	501.23	483.22	484.21	251.12	501.23	G	800.50	782.49	783.47	400.75	800.50	8
6	572.27	554.25	555.24	286.63	572.27	A	743.47	725.47	726.45	372.24	743.48	7
7	685.34	667.34	668.32	343.18	685.35	L	672.44	654.43	655.41	336.72	672.44	6
8	786.41	768.37	769.38	393.70	786.40	T	559.36	541.35	542.33	280.18	559.36	5
9	843.42	825.41	826.39	422.21	843.42	G	458.31	440.30	441.28	229.65	458.31	4
10	956.50	938.49	939.48	478.75	956.50	L	401.29	383.28	384.26	201.14	401.29	3
11	1069.59	1051.58	1052.56	535.29	1069.59	I(sub T)	288.20	270.19	271.18	144.60	288.20	2
12						R	175.12	157.11	158.09	88.06	175.12	1

Supplementary Figure 4. NELSGALTGLIR in Sample 2



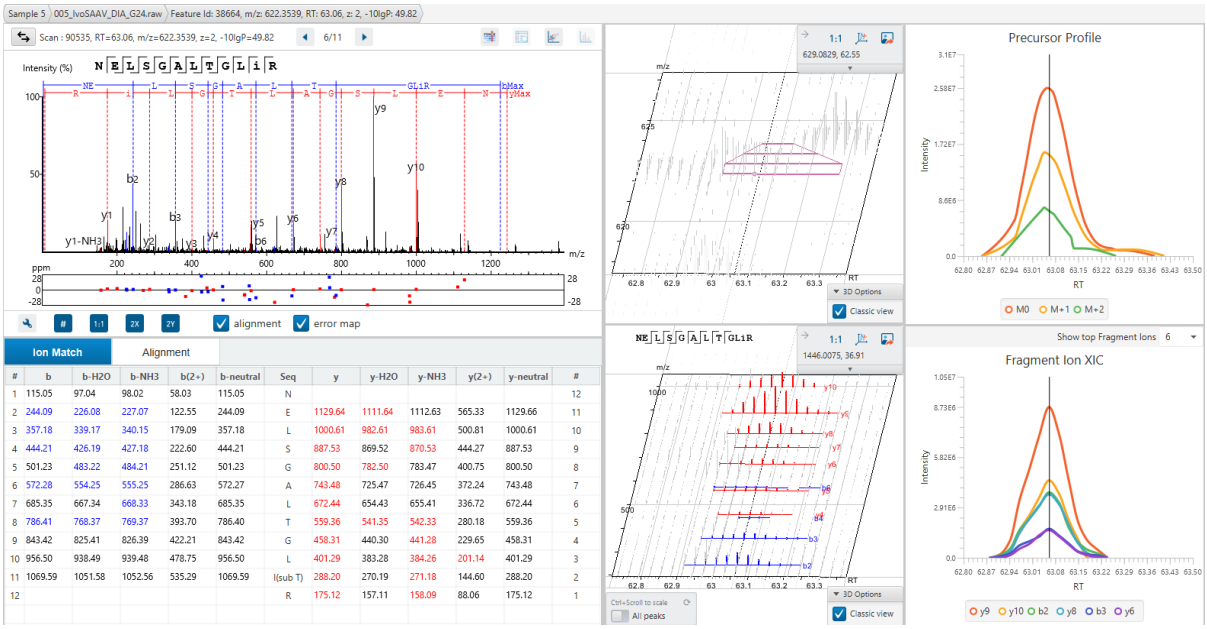
#	b	b-H2O	b-NH3	b(2+)	b-neutral	Seq	y	y-H2O	y-NH3	y(2+)	y-neutral	#
1	115.05	97.04	98.02	58.03	115.05	N						12
2	244.09	226.08	227.07	122.55	244.09	E	1129.63	1111.63	1112.63	565.33	1129.66	11
3	357.18	339.17	340.15	179.09	357.18	L	1000.61	982.60	983.60	500.81	1000.61	10
4	444.21	426.20	427.18	222.60	444.21	S	887.53	869.52	870.52	444.27	887.53	9
5	501.23	483.22	484.20	251.12	501.23	G	800.50	782.50	783.47	400.75	800.50	8
6	572.27	554.26	555.24	286.63	572.27	A	743.48	725.47	726.45	372.23	743.48	7
7	685.35	667.34	668.32	343.18	685.35	L	672.44	654.43	655.41	336.72	672.44	6
8	786.41	768.37	769.39	393.70	786.40	T	559.36	541.35	542.33	280.18	559.36	5
9	843.42	825.41	826.39	422.21	843.42	G	458.31	440.30	441.28	229.65	458.31	4
10	956.50	938.49	939.48	478.75	956.50	L	401.29	383.28	384.26	201.14	401.29	3
11	1069.59	1051.58	1052.56	535.29	1069.59	I(sub T)	288.20	270.18	271.18	144.60	288.20	2
12						R	175.12	157.11	158.09	88.06	175.12	1

Supplementary Figure 5. NELSGALTGLIR in Sample 3



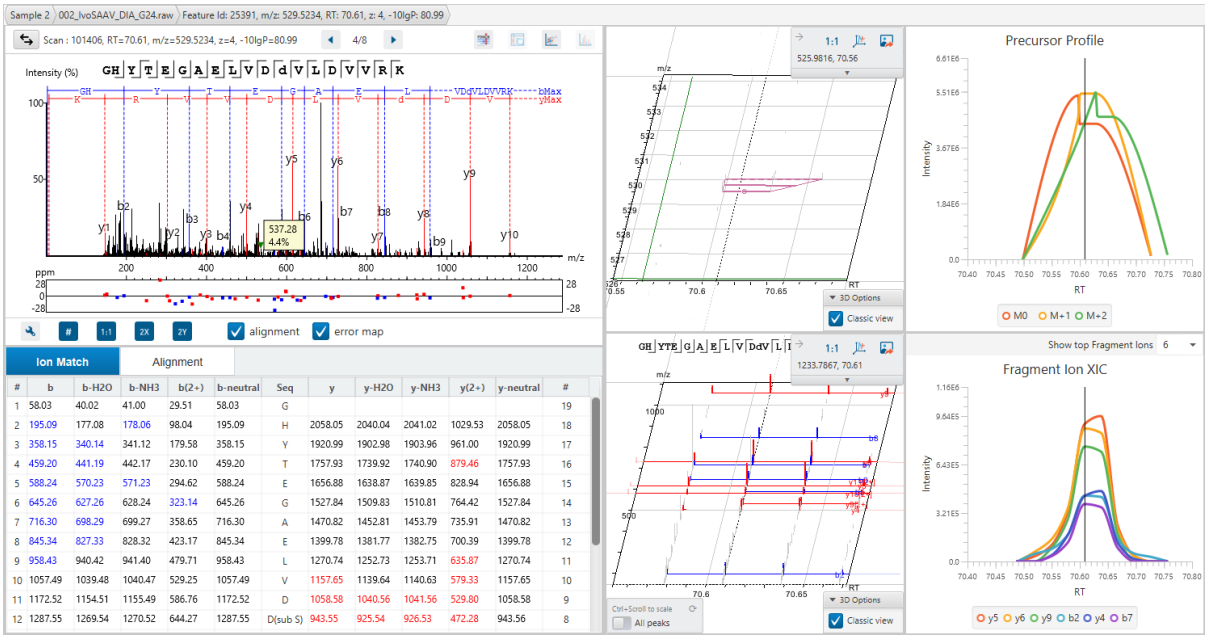
#	b	b-H2O	b-NH3	b(2+)	b-neutral	Seq	y	y-H2O	y-NH3	y(2+)	y-neutral	#
1	115.05	97.04	98.02	58.03	115.05	N						12
2	244.09	226.08	227.07	122.55	244.09	E	1129.64	1111.62	1112.63	565.33	1129.66	11
3	357.18	339.17	340.15	179.09	357.18	L	1000.61	982.61	983.60	500.81	1000.61	10
4	444.21	426.20	427.18	222.60	444.21	S	887.53	869.53	870.52	444.27	887.53	9
5	501.23	483.22	484.21	251.12	501.23	G	800.50	782.50	783.47	400.75	800.50	8
6	572.27	554.25	555.25	286.63	572.27	A	743.48	725.47	726.45	372.23	743.48	7
7	685.35	667.34	668.32	343.18	685.35	L	672.44	654.43	655.41	336.72	672.44	6
8	786.41	768.38	769.38	393.70	786.40	T	559.36	541.35	542.33	280.18	559.36	5
9	843.42	825.41	826.39	422.21	843.42	G	458.31	440.30	441.28	229.65	458.31	4
10	956.50	938.49	939.48	478.75	956.50	L	401.29	383.28	384.26	201.14	401.29	3
11	1069.59	1051.58	1052.56	535.29	1069.59	I(sub T)	288.20	270.19	271.18	144.60	288.20	2
12						R	175.12	157.11	158.09	88.06	175.12	1

Supplementary Figure 6. NELSGALTGLIR in Sample 4



#	b	b-H2O	b-NH3	b(2+)	b-neutral	Seq	y	y-H2O	y-NH3	y(2+)	y-neutral	#
1	115.05	97.04	98.02	58.03	115.05	N						12
2	244.09	226.08	227.07	122.55	244.09	E	1129.64	1111.64	1112.63	565.33	1129.66	11
3	357.18	339.17	340.15	179.09	357.18	L	1000.61	982.61	983.61	500.81	1000.61	10
4	444.21	426.19	427.18	222.60	444.21	S	887.53	869.52	870.53	444.27	887.53	9
5	501.23	483.22	484.21	251.12	501.23	G	800.50	782.50	783.47	400.75	800.50	8
6	572.28	554.25	555.25	286.63	572.27	A	743.48	725.47	726.45	372.24	743.48	7
7	685.35	667.34	668.33	343.18	685.35	L	672.44	654.43	655.41	336.72	672.44	6
8	786.41	768.37	769.37	393.70	786.40	T	559.36	541.35	542.33	280.18	559.36	5
9	843.42	825.41	826.39	422.21	843.42	G	458.31	440.30	441.28	229.65	458.31	4
10	956.50	938.49	939.48	478.75	956.50	L	401.29	383.28	384.26	201.14	401.29	3
11	1069.59	1051.58	1052.56	535.29	1069.59	I(sub T)	288.20	270.19	271.18	144.60	288.20	2
12						R	175.12	157.11	158.09	88.06	175.12	1

Supplementary Figure 7. NELSGALTGLIR in Sample 5



#	b	b-H2O	b-NH3	b(2+)	b-neutral	Seq	y	y-H2O	y-NH3	y(2+)	y-neutral	#
1	58.03	40.02	41.00	29.51	58.03	G						19
2	195.09	177.08	178.06	98.04	195.09	H	2058.05	2040.04	2041.02	1029.53	2058.05	18
3	358.15	340.14	341.12	179.58	358.15	Y	1920.99	1902.98	1903.96	961.00	1920.99	17
4	459.20	441.19	442.17	230.10	459.20	T	1757.93	1739.92	1740.90	879.46	1757.93	16
5	588.24	570.23	571.23	294.62	588.24	E	1656.88	1638.87	1639.85	828.94	1656.88	15
6	645.26	627.26	628.24	323.14	645.26	G	1527.84	1509.83	1510.81	764.42	1527.84	14
7	716.30	698.29	699.27	358.65	716.30	A	1470.82	1452.81	1453.79	735.91	1470.82	13
8	845.34	827.33	828.32	423.17	845.34	E	1399.78	1381.77	1382.75	700.39	1399.78	12
9	958.43	940.42	941.40	479.71	958.43	L	1270.74	1252.73	1253.71	635.87	1270.74	11
10	1057.49	1039.48	1040.47	529.25	1057.49	V	1157.65	1139.64	1140.63	579.33	1157.65	10
11	1172.52	1154.51	1155.49	586.76	1172.52	D	1058.58	1040.56	1041.56	529.80	1058.58	9
12	1287.55	1269.54	1270.52	644.27	1287.55	D(sub S)	943.55	925.54	926.53	472.28	943.56	8

Supplementary Figure 8. A low abundant peptide variant reported by DIAVariant.