Platinum® Next-Generation Protein Sequencing™ Advanced Data Analysis Technical Note

Next-Generation Protein Sequencing™ (NGPS) on Platinum® enables protein identification and characterization with single-molecule and single amino acid resolution in a simple workflow using a benchtop instrument. Platinum Analysis Software workflows make peptide alignment and protein mapping easy to interpret without the need for advanced expertise. The output files from Platinum Analysis Software provide researchers with the opportunity to perform advanced analysis on NGPS data including analysis of amino acid variants and modifications based on kinetic signatures. This technical note provides guidance on interpreting results from Primary Analysis and Peptide Alignment Workflows and performing advanced analysis on NGPS data generated by Platinum.

Q-SI TECHNOLOGY

Quantum-Si's benchtop Platinum™ instrument enables protein sequencing from biological samples in a simple user-friendly workflow. Our technology utilizes dye-tagged N-terminal amino acid recognizers and semiconductor chip technology to detect the binding characteristics and binding order of N-terminal amino acids, resulting in unique kinetic signatures that can be used to differentiate and identify amino acid residues and PTMs. A more detailed overview of the workflow and technology can be found in our Science Paper.

PLATINUM NEXT-GENERATION PROTEIN SEQUENCING

Sequencing proteins on Platinum begins by digesting proteins into functionalized peptides that are then immobilized into nanoscale reaction chambers on a semiconductor chip. Fluorescently labeled N-terminal amino acid (NAA) recognizers and aminopeptidases are then added to the chip. The sequencing process commences as the recognizers interact with each NAA and generate a fluorescent signal from which the binding kinetics characteristic of each amino acid are extracted. This signal is collected by individual sensors associated with each reaction chamber. Specifically, as the recognizers repeatedly associate and dissociate with the NAAs, a distinct series of pulses, termed a recognition segment (RS), is generated for each recognized NAA with charac-

teristic fluorescence and kinetic properties. Aminopeptidases sequentially cleave NAA, exposing the next NAA for recognition until the entire peptide is sequenced. The recognizers are distinguished by the fluorescent intensity and lifetime of each binding event. The temporal order of NAA recognition and the associated kinetic properties (interpulse duration and pulse duration) of each RS, referred to as a kinetic signature, are used to align to amino acid, peptide, and protein sequences.

PRIMARY ANALYSIS WORKFLOW

Primary Analysis performs signal processing on the raw fluorescent intensity and lifetime data collected by Platinum. The main output of Primary Analysis is reads, which are sequences of recognition segments associated with each recognizer. Reads are characterized by the recognizer read length or the number of distinguishable recognizers per read. Primary Analysis also produces plots that are useful for run characterization, such as chip loading percentage and activity over time.

SETTING UP THE PRIMARY ANALYSIS WORKFLOW

Instructions explaining how to set up and view Primary Analysis can be found in the Platinum Analysis Software User Manual (Document # PTL-0006).

UNDERSTANDING THE RESULTS FROM THE PRIMARY ANALYSIS WORKFLOW

The Summary page displays key information about the data collected from apertures in a run, including:

1. High-Quality (HQ) Reads: Total number of reads with recognizer read lengths ≥ 4 and unique active recognizers ≥ 3, which are then considered for downstream analysis workflows, such as Peptide Alignment.
2. Loading: Percentage of analyzed apertures that were loaded. The loading percentage includes both single-loaded and multi-loaded apertures.

The Plots and the Recognizer Reads pages display key information about apertures and recognizer activity across the chip:

1. Loading Heatmap **(Figure 1A)**: XY image indicating the position of the loaded apertures and total loading percentage. The dark blue regions are excluded from usable apertures. Loaded apertures should be homogenously distributed across the chip.
2. Recognizer Activity Over Time **(Figure 1B)**: Time course plot indicating the number of analyzed apertures in which a specific recognizer was detected throughout the run. This plot will vary as a function of peptide composition of the sample. For example, if the sample contains a single peptide, the plot will display increased activity for each recognizer as each residue in the peptide

is cleaved during the run. If the sample contains a diverse population of peptides, the recognizer activity will appear more uniform over time.

3. Recognizer Read Lengths Heatmap **(Figure 1C)**: XY image indicating the recognizer read lengths of each aperture across all analyzed apertures. Read lengths should be homogenously distributed across the chip.

4. Recognizer Read Lengths Histogram **(Figure 1D)**: Bar chart providing the total number of reads for each read length across all analyzed apertures.

5. Recognizer Reads **(Figure 2)**: Bar chart summarizing the total number of reads for each of the RS motifs identified. The RS motifs are sorted from the highest to lowest number of reads. This plot provides insights into overall chip recognizer activity.

FIGURE 1

A) Loading Heatmap – an XY image of the chip indicating the position of loaded apertures. B) Recognizer Activity Over Time, a bar chart representing the number of apertures over the course of a run in which each recognizer is active. The upper figure shows an example of a single peptide (ELRAQFAYPDDDK) with distinct timing of the appearance of each Recognizer as function of the aminopeptidase cutting during the run. The lower figure is a plot from a mixture of 10 proteins which shows a more uniform appearance of the recognizers due to the diversity of the amino acid composition of the sample. C) Recognizer Read Lengths Heatmap – an XY image of the chip indicating the read length of each analyzed aperture. D) Recognizer Read Lengths Histogram, a bar chart representing the numbers of reads for each read length

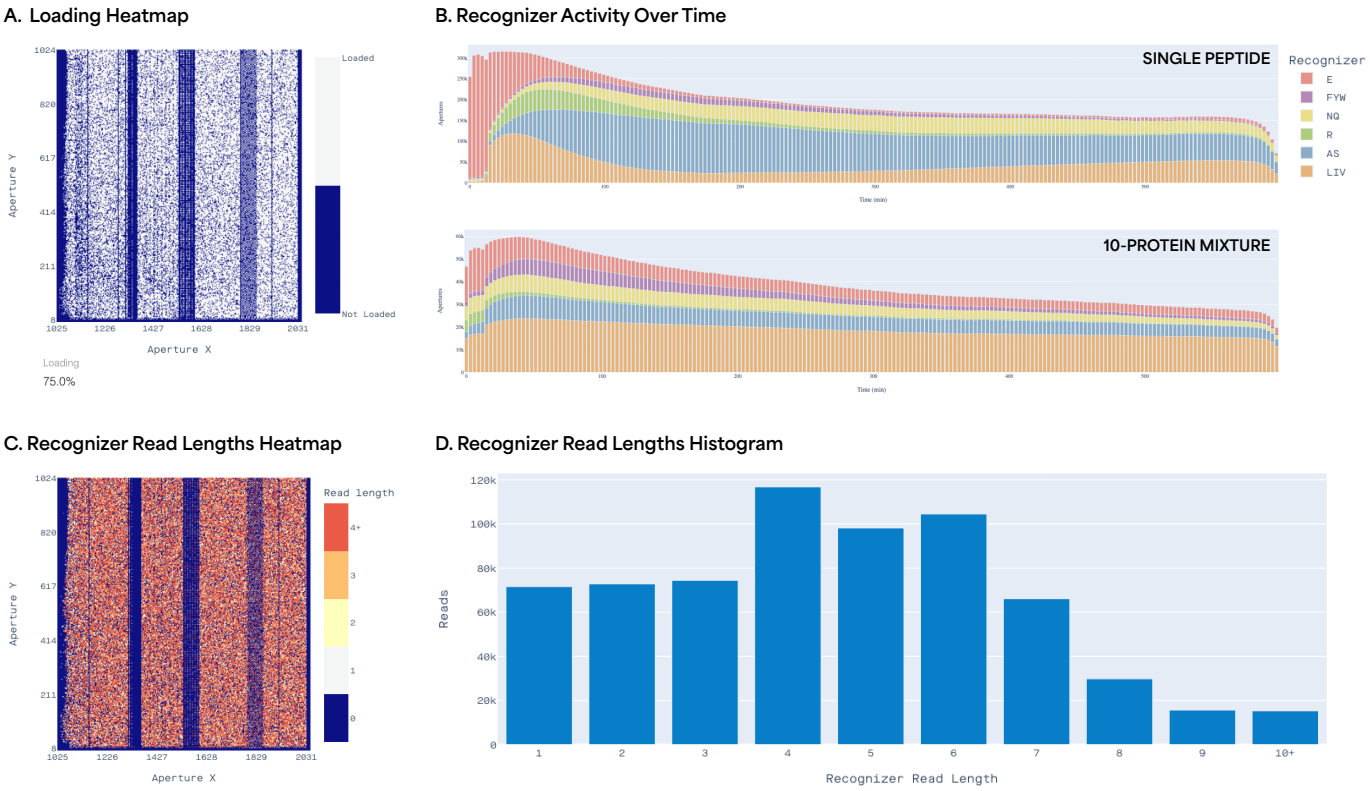
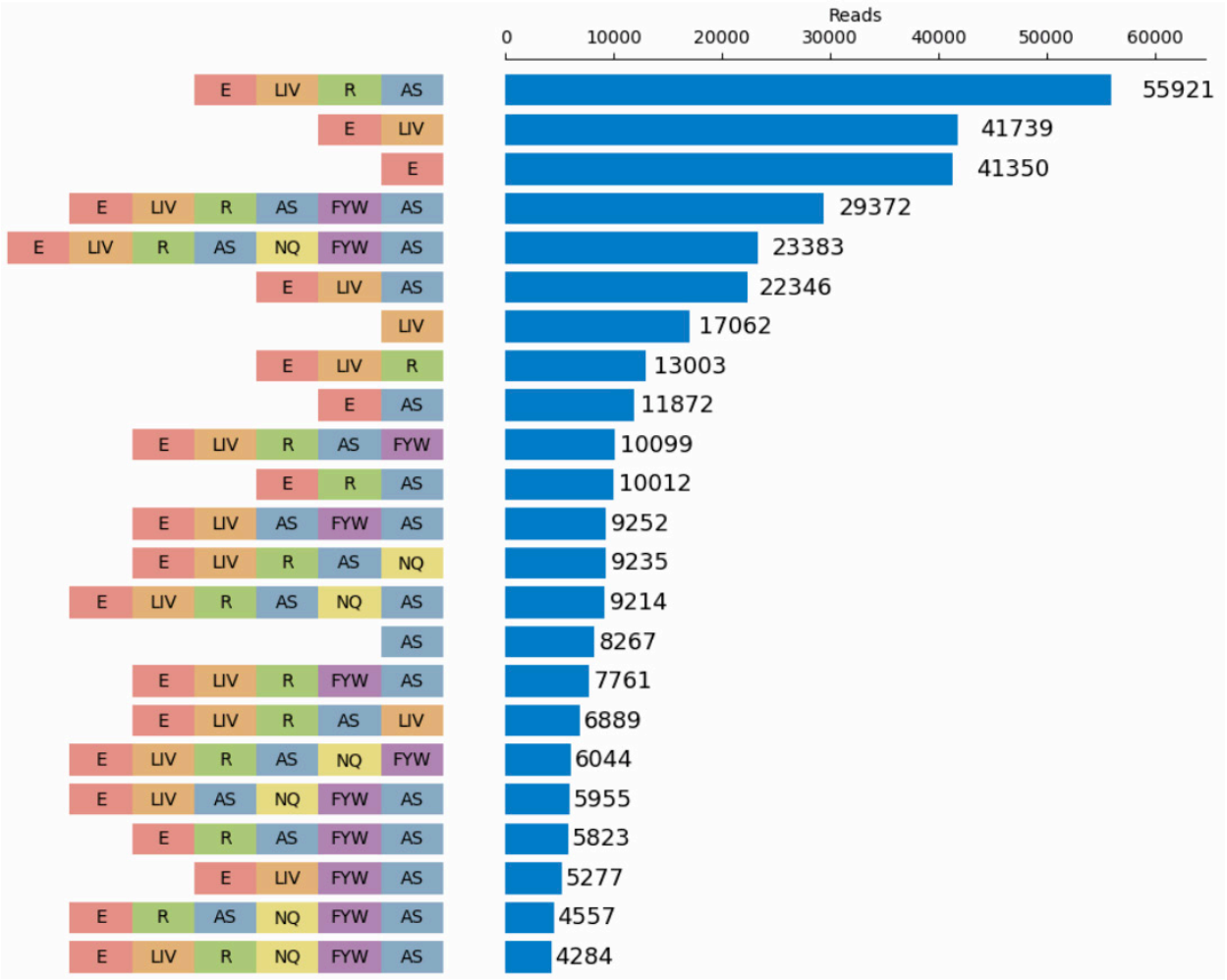


FIGURE 2: RECOGNIZER READS

A bar chart indicating the number of reads detected for all observed recognition segment motifs. The RS motifs are ordered form the most abundant to the least. Please note that this figure is cropped and does not display all the recognizer reads from the example run.



DOWNLOADABLE FILES FROM PRIMARY ANALYSIS WORKFLOW

The *recognition_segments.csv* file is available for advanced analysis of protein sequencing run data. This file contains a summary of the data used to characterize all apertures containing recognition segments (RS) that pass quality filters. It can be used to determine patterns of recognizer motifs prior to alignment workflows.

The *recognition_segments.csv* file contains the following information **(Table 1)**:

1. aperture_index: Unique numerical identifier for each aperture on the chip containing a quality filter passed RS.

2. rs_index: Numerical identifier for each RS detected in a specific aperture, starting at 0 for the first

detected RS and increasing by 1 with each new detection. Each recognition segment will have a consistent pulse duration (PD), interpulse duration (IPD), dye lifetime, and intensity properties. Significant changes in those properties between two pulsing events will result in the detection of a new RS. A recognition segment that cannot be accurately discriminated by the above properties will be filtered and excluded from the rs_index annotation. Apertures with filtered RS will have gaps in the numerical rs_index order.

3. recognizer: Identity of the recognizer associated with each RS, determined by fluorescence lifetime and intensity.
4. start_s: Start time of each RS in seconds from the start of the run.
5. end_s: End time of the same RS in seconds from the start of the run.
6. pd_s: Calculated mean PD of the specific RS in seconds, representing the average residence time of all the binding events between the recognizer and the N-terminal amino acid (NAA). Each RS may contain 10s or 100s of individual pulses that are used to calculate the mean pulse duration of the RS. The number of individual pulses is determined by the duration of the RS as well as the overall binding kinetics of the interaction.
7. ipd_s: Calculated mean IPD of the specific RS in seconds, representing the average time between two individual binding events within a specific RS.

TABLE 1

Aperture level view of recognition segments and kinetics information for Aperture 228714.

aperture_index	rs_index	recognizer	start_s	end_s	pd_s	ipd_s
228714	0	E	60.8574	869.2888	1.1259	5.9971
228714	1	LIV	913.0413	1700.8267	0.3845	5.6519
228714	2	R	1708.749	2465.3855	0.3792	2.2162
228714	3	AS	2468.9266	9036.9662	0.6313	5.8284
228714	4	NQ	9048.0093	9259.81	3.621	8.3891
228714	5	FYW	9260.5902	9699.0156	2.0292	3.7559
228714	6	AS	11934.2353	12455.4845	0.6722	26.4316
228714	7	FYW	12538.5483	12857.4795	0.3001	20.92

Aperture number 228714 in **Table 1** will be used as an example on how to interpret data in the *recognition_segments.csv* file. In this aperture, 8 RSs were detected. The first RS started around 1 minute into the run, and the final RS ended approximately ~4 hours after the start of the run.

Breaking down the detected RSs in sequence order:

1. The 1st RS (rs_index 0) corresponds to the E recognizer.
2. The 2nd RS (rs_index 1) corresponds to the LIV recognizer.
3. The 3rd RS (rs_index 2) corresponds to the R recognizer.
4. The 4th RSs (rs_index 3) correspond to the AS recognizer.
5. The 5th RS (rs_index 4) corresponds to the NQ recognizer.
6. The 6th RS (rs_index 5) corresponds to the FYW recognizer.
7. The 7th RS (rs_index 6) corresponds to the AS recognizer.
8. The 8th RS (rs_index 7) corresponds to the FYW recognizer.

The probable recognizer order for this aperture is E-LIV-R-AS-NQ-FYW-AS-FYW resulting in a recognizer read length of 8. The PD and IPD values calculated in the Primary Analysis workflow will be used to align reads to reference peptide and protein sequences in the Peptide Alignment Workflow.

PEPTIDE ALIGNMENT WORKFLOW

Peptide Alignment is an analysis workflow designed to align reads identified in Primary Analysis to reference peptide and protein sequences. Once the Primary Analysis has completed, the secondary Peptide Alignment workflow can be executed. Protein or peptide reference sequences are provided by the user as a FASTA file, and the Peptide Alignment workflow uses empirical PD values and the RS order of the reads from Primary Analysis to align reads to the predicted PD values and RS order of the reference sequence (based on a database of expected kinetic signatures).

SETTING UP THE PEPTIDE ALIGNMENT WORKFLOW

Instructions detailing how to set up and view Peptide Alignment analysis can be found in the Platinum Analysis Software User Manual (Document # PTL-0006).

UNDERSTANDING THE RESULTS FROM THE PEPTIDE ALIGNMENT WORKFLOW

The Summary page displays the following information:

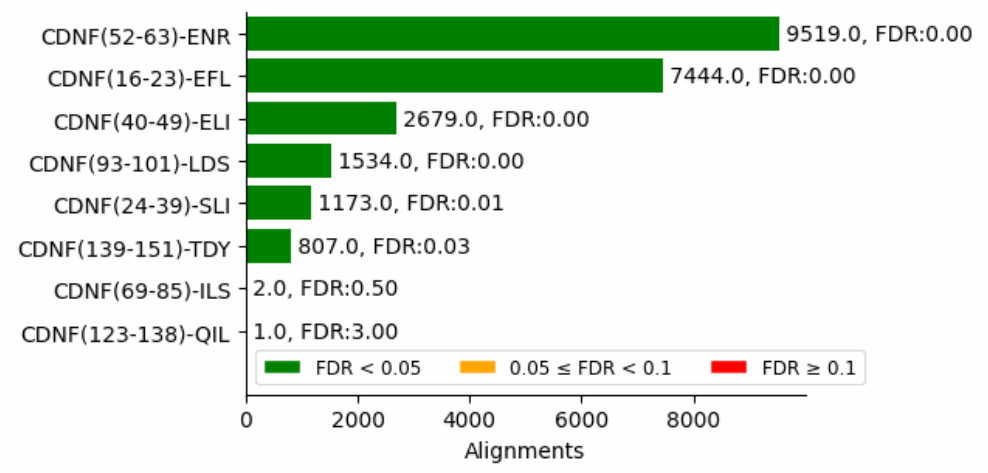
1. Alignments: Total number of alignments that pass strict filters with a minimum alignment score of 4.0. This value represents the goodness of fit of the observed RS order and PD data for each read from Primary Analysis to the reference sequence. For more information on the alignment score, see the section in this document on the *alignments.csv* file.
2. Peptides Identified: Total number of peptides identified.

The Alignment tab will list the total number of alignments for each detected peptide or protein. The

peptide-specific alignment plot includes the number of alignments and the False Discovery Rate (FDR) for each identified peptide, sorted from highest to lowest alignments. FDR is calculated using a decoy generation method adapted from methods used in peptide identification by mass spectrometry; peptides with FDR of <5% are colored in green, between 5% to 10% are colored in orange, and >10% are colored in red. This approach is similarly used to display alignments and FDR from multi-protein samples **(Figure 3)**.

FIGURE 3: VIEW OF ALIGNMENT TAB FROM PEPTIDE ALIGNMENT WORKFLOW FOR A SINGLE PROTEIN.

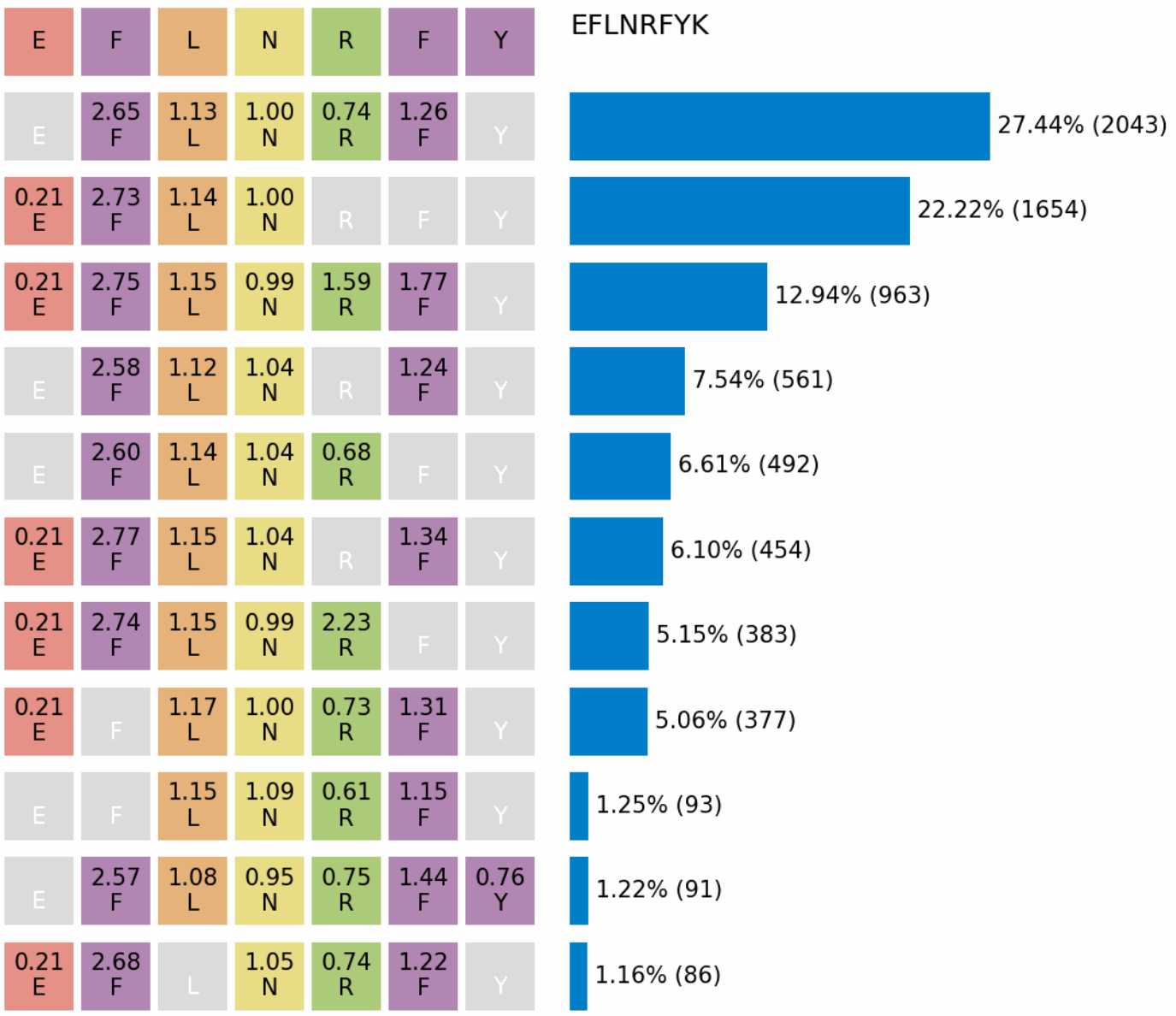
The bar charts represent the number of alignments for peptides along with peptide level FDR.



The Profiles tab **(Figure 4)** displays the peptides identified with at least 20 alignments. The color-filled box plot shows the RS pattern with the mean PD of each RS. The side bar chart displays, for each peptide, the percentage of the total alignments that match the RS pattern, along with the number of alignments in parentheses, sorted from highest to lowest abundance. Note that only aligned apertures with an alignment score of at least 4.0 are included, potentially reducing the number alignments compared to the reads from the Primary Analysis workflow.

FIGURE 4: VIEW OF PROFILES TAB FROM PEPTIDE ALIGNMENT WORKFLOW.

Top row: Reference sequence. Other rows: Observed patterns with mean PD (grey boxes indicate undetected). Side bar chart: Displays number of alignments and the percentage of total analyzed apertures that match corresponding patterns. Note: This figure has been cropped and does not display all possible peptide profiles from the run.



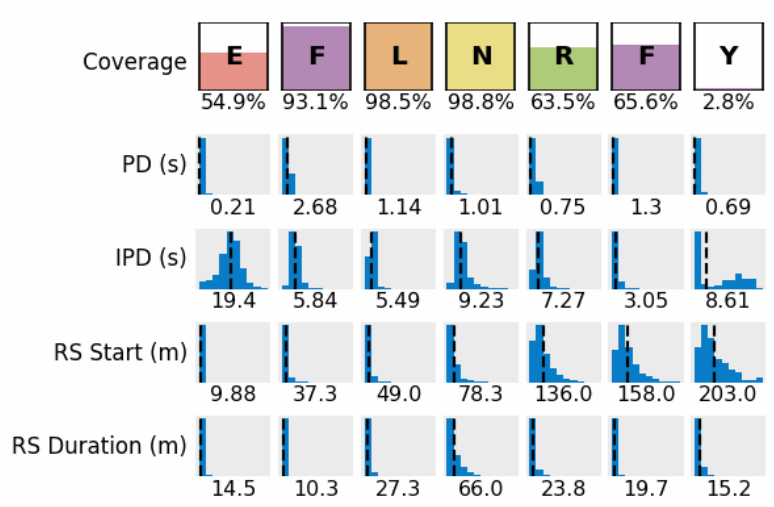
The Peptides tab displays the protein sequence of the reference sequence and the number of peptide alignments for each peptide. For peptides with at least 20 total alignments, the per-position sequencing coverage and global kinetic signature of all identified peptides are displayed **(Figure 5)**. The plots represent the cumulative data for each aligned motif from the Profiles tab, providing summary statistics for all aligned apertures to a given reference sequence.

- Coverage: Represents the percent of alignments with each RS by fill color height of each RS box.

- PD (s): The pulse duration reported as the median of the mean PD for all RSs aligned to the given position, displayed as boxed histogram plots with values in seconds.
- IPD (s): The interpulse duration reported as the median of the mean IPD for all RSs aligned to the given position, displayed as boxed histogram plots with values in seconds.
- RS Start (m): Boxed histogram plots of the mean start times for all RSs aligned to the given position, with time in minutes displayed below the plot.
- RS Duration (m): Boxed histogram plots of the mean durations for all RSs aligned to the given position, with time in minutes displayed below the plot.

FIGURE 5: EXAMPLE OF THE PEPTIDE TAB DATA.

View of Peptides tab, showing the sequence of the observed peptide and the corresponding coverage and kinetics information.



DOWNLOADABLE FILES FROM PEPTIDE ALIGNMENT WORKFLOW

For advanced data analysis, the *peptide_reference.csv* and *alignments.csv* files from the Peptide Alignment workflow are available for download. The files contain peptide and protein alignments and data from individual apertures. The information contained in these files, when combined with the *recognition_segments.csv* file from the Primary Analysis workflow, can be used to reconstruct the full kinetic sequencing data set for each available aperture, excluding data that did not pass quality filters. In addition, the data contained in these files can enable the user, if desired, to develop their own custom analysis scripts to further analyze the data.

Peptide_reference.csv

The *peptide_reference.csv* file provides summary information in a column-based format from the Peptide Alignment analysis workflow. It includes details such as protein name, peptide name, ref-

erence sequence, total length of the reference sequence, FDR, and the observed total number of alignments. Below is an example of the output from a *peptide_reference.csv* file for a single protein run **(Table 2)**.

1. Protein: The name of the protein from the user supplied reference fasta file for which the peptide was aligned, if applicable.
2. Peptide: The aligned reference peptides sequence displayed as Protein Name-(amino acid positions)-First three amino acids of the peptide sequence as determined by *in silico* LysC digestion.
3. Sequence: The full amino acid sequence of the aligned peptide as determined by *in silico* LysC digestion of the reference sequence.
4. Length: The number of amino acids in the peptide.
5. FDR: The calculated false discovery rate of the aligned peptide.
6. Alignments: The total number of reads that were aligned to each peptide.

TABLE 2: EXAMPLE OF PEPTIDE REFERENCE TABLE.

Protein	Peptide	Sequence	Length	FDR	Alignments
CDNF	CDNF(52-63)-ENR	ENRLCYYLGATK	12	0.0002	9519
CDNF	CDNF(16-23)-EFL	EFLNRFYK	8	0.0004	7444
CDNF	CDNF(40-49)-ELI	ELISFCLDTK	10	0.0011	2679
CDNF	CDNF(93-101)-LDS	LDSQICELK	9	0.0013	1534
CDNF	CDNF(24-39)-SLI	SLIDRGVNFSLDTIEK	16	0.006	1173
CDNF	CDNF(139-151)-TDY	TDYVNLIQELAPK	13	0.031	807
CDNF	CDNF(69-85)-ILS	ILSEVTRPMSVHMPAMK	17	0.5	2
CDNF	CDNF(123-138)-QIL	QILHSWGEECRACAEK	16	3	1

Alignments.csv

The *alignments.csv* file, in a column-based format, provides information from all aligned reads analyzed during the Peptide Alignment workflow **(Table 3)**.

1. aperture_index: Unique numerical identifier to enable association of the read data to each aperture on the chip. Corresponds to aperture_index in the *recognition_segments.csv* file from Primary Analysis.
2. peptide: Aligned reference peptide sequences displayed as Protein Name-(amino acid positions)-First three amino acids of the peptide sequence.
3. rs_indices: List of RS identified and the order in which they were aligned to the reference se-

quence. The rs_indices are based on rs_index in the *recognition_segments.csv* file from Primary Analysis.

4. ref_indices: List of the positions within the reference sequence to which an RS is aligned. Each ref_index number corresponds to a specific amino acid within the reference peptide sequence and uses a similar 0-based number scheme as the rs_index with the first amino acid in the sequence assigned a value of 0 and the subsequent amino acids incrementing +1.

5. aln_score: A metric indicating the agreement between the kinetic data of an observed read and the expected kinetic signature of a given reference peptide. The score is based on the presence of expected RS's in the correct order and the agreement between the observed PD value for each RS and the expected PD value based on our kinetic database. A minimum alignment score of 4.0 is required for an aperture to be aligned to the reference sequence.

TABLE 3: EXAMPLE OF ALIGNMENTS TABLE FOR CDNF PROTEIN.

For a protein sample alignment file, you will see all the different peptides from a specific protein listed under peptide name.

aperture_index	peptide	rs_indices	ref_indices	aln_score
598556	CDNF(52-63)-ENR	[0 1 2 3 4 5 6]	[0, 1, 2, 3, 5, 6, 7]	8.2366
637445	CDNF(16-23)-EFL	[0 1 2 3 4 5]	[0, 1, 2, 3, 4, 5]	8.212
1362493	CDNF(93-101)-LDS	[0 1 2 3 4 5 6]	[0, 2, 3, 4, 6, 6, 7]	7.5689
377482	CDNF(139-151)-TDY	[0 5 6 7 8 9 10 11 12]	[2, 4, 5, 6, 6, 7, 9, 9, 10]	7.0486

Understanding the Assignment of Primary Analysis Data to the Reference Sequence

The alignment of aperture data to a peptide always evaluates the goodness of fit for both the order of RS appearance and the observed PD of each RS to a kinetic database model of all possible reference peptides. Therefore, to interpret and extract kinetic data from the downloadable files, it is crucial to understand the relationship between the reference sequence and the observed kinetic data from the Primary Analysis. The protein or peptide reference sequence is provided by the user as a fasta file when setting up the analysis. After digestion with LysC, each amino acid in the peptide reference sequence is assigned a reference index (ref_index) corresponding to its position in the peptide chain, excluding internal prolines, any residues downstream of internal prolines and the terminal lysine residue. For example, in the case of the CDNF-EFLNRFYK peptide, the first amino acid (E) has a ref_index of 0, and the last amino acid (Y) has a ref_index of 6 **(Figure 6)**.

Once the reference positions are assigned a ref_index, one must determine whether the RS data from the primary analysis workflow contains matches, duplications, or deletions to the ref_index positions in the provided reference sequence. In a simple case of a direct match of the primary RS data to the reference sequence, such as aperture 637445 **(Figure 6A)**, the rs_indices will align in a direct 1:1 relationship to the ref_indices; that is rs_index 0 aligns to ref_index 0, rs_index 1 aligns to ref_index 1, and so on until all rs_indices are aligned to the ref_indices. In the event that there is a duplication in which multiple rs_indices are mapped to the same ref_index position, such as ref_index 1 shown in Aperture

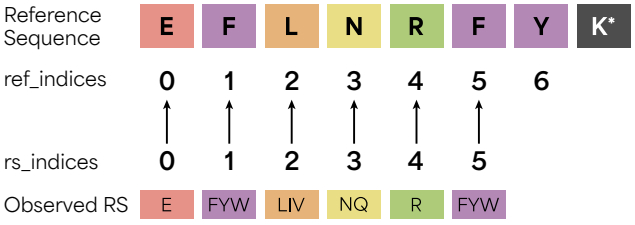
406302 **(Figure 6B)**, the data must be aligned differently. In this case, the data maintains the same 1:1 relationship, however all RS's that map to the same ref_index position are aggregated into a single RS. The resulting combined RS will have a total duration that spans all aggregated rs_indices, with the PD and IPD calculated as the average of all aggregated rs_indices. Finally, if a ref_index position is not detected, such as the deletion of the ref_index 0 shown in Aperture 94594 **(Figure 6C)**, the data must be aligned differently to account for this change. As in the previous examples, the 1:1 relationship is maintained, but the rs_indices are shifted to the appropriate ref_index positions based on the fit of the primary analysis data to the reference sequence.

FIGURE 6: EXAMPLES OF ALIGNING RECOGNITION SEGMENTS WITH EXACT MATCHES, DUPLICATIONS, OR DELETIONS TO RESIDUES IN THE REFERENCE SEQUENCE.

Each example is associated with a table showing the assignment of rs_indices to ref_indices from the Peptide Alignment workflow for each aperture. A) Aperture 637445 shows an exact match with all rs_indices aligned to ref_indices in a direct, 1:1 relationship. B) Aperture 406302 contains a duplication of ref_index 1, which is resolved by combining rs_indices 1 and 2 into a single RS aligning to the F residue at ref_index position 1. C) Aperture 94594 contains a deletion of ref_index 0, resulting in the shift of the rs_indices by one position to account for the inability to detect the E residue in the aperture.

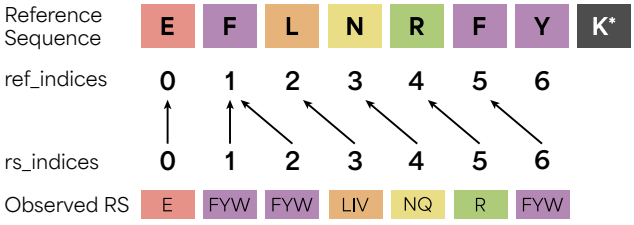
A. Aperture 637445

aperture_index	peptide	rs_indices	ref_indices
637445	CDNF(16-23)-EFL	[0 1 2 3 4 5]	[0, 1, 2, 3, 4, 5]



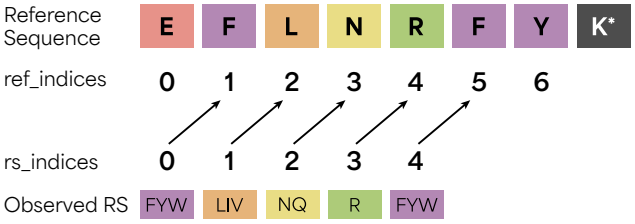
B. Aperture 406302

aperture_index	peptide	rs_indices	ref_indices
406302	CDNF(16-23)-EFL	[0 1 2 3 4 5 6]	[0, 1, 1, 2, 3, 4, 5]



C. Aperture 94594

aperture_index	peptide	rs_indices	ref_indices
94594	CDNF(16-23)-EFL	[0 1 2 3 4]	[1, 2, 3, 4, 5]



Reconstructing alignments from kinetic data

Once the relationship between a reference sequence and RSs has been established, the full kinetic trace data for each aperture can be reconstructed by examining the alignment data from the Peptide

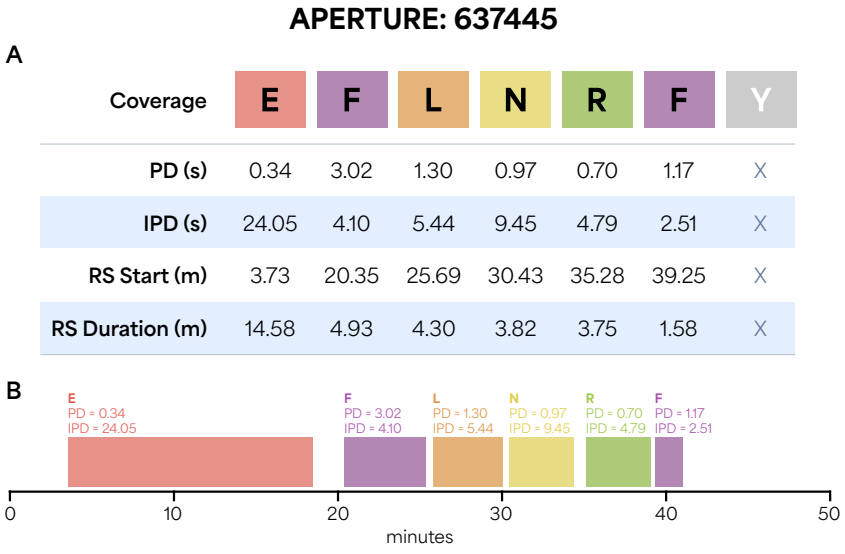
Alignment workflow in combination with the kinetic data generated in the Primary Analysis workflow. By breaking down the kinetic data for each detected RS in the appropriate sequence order, we can better understand how each aperture is aligned to the assigned peptide based on the overall kinetic data.

aperture_index	rs_index	recognizer	start_s	end_s	pd_s	ipd_s
637445	0	E	223.8908	1099.2857	0.3418	24.0564
637445	1	FYW	1221.3752	1517.2351	3.0276	4.1029
637445	2	LIV	1541.4849	1799.7697	1.3	5.4419
637445	3	NQ	1826.3004	2055.7134	0.9734	9.4564
637445	4	R	2116.9983	2342.4497	0.7053	4.7914
637445	5	FYW	2361.5974	2456.9161	1.1741	2.515

Again, using aperture 637445 as an example, the kinetic data **(Table 4)** from the *recognition_segments.csv* file can be associated to each residue in the reference sequence to produce the specific kinetic trace data for this aperture. The fully reconstructed kinetic data displaying the PD, IPD, RS start time and duration is shown for aperture 637445 in **Figure 7**. The process of reconstructing the kinetic trace data can be repeated for all aligned apertures within the downloadable files.

FIGURE 7: GRAPHICAL SUMMARY OF THE KINETIC DATA OF APERTURE 637445.

A) The aperture specific kinetic data was reconstructed from the information in both the *alignments.csv* and *recognition_segments.csv* files. B) Visual representation of the real time sequencing of the EFLNRFYK peptide. Colored segments represent the duration of each RS before cleavage by aminopeptidases. The calculated PD and IPD for each RS is reported for the colored segment.



SUMMARY

This technical note demonstrates the power of using kinetic signatures to identify and characterize peptides and proteins. Advanced analysis using the data from Primary Analysis and Peptide Alignment workflows enables researchers to explore how variations in protein sequencing data can be correlated to biological function. For more information and guidance related to other applications, contact support@quantum-si.com.