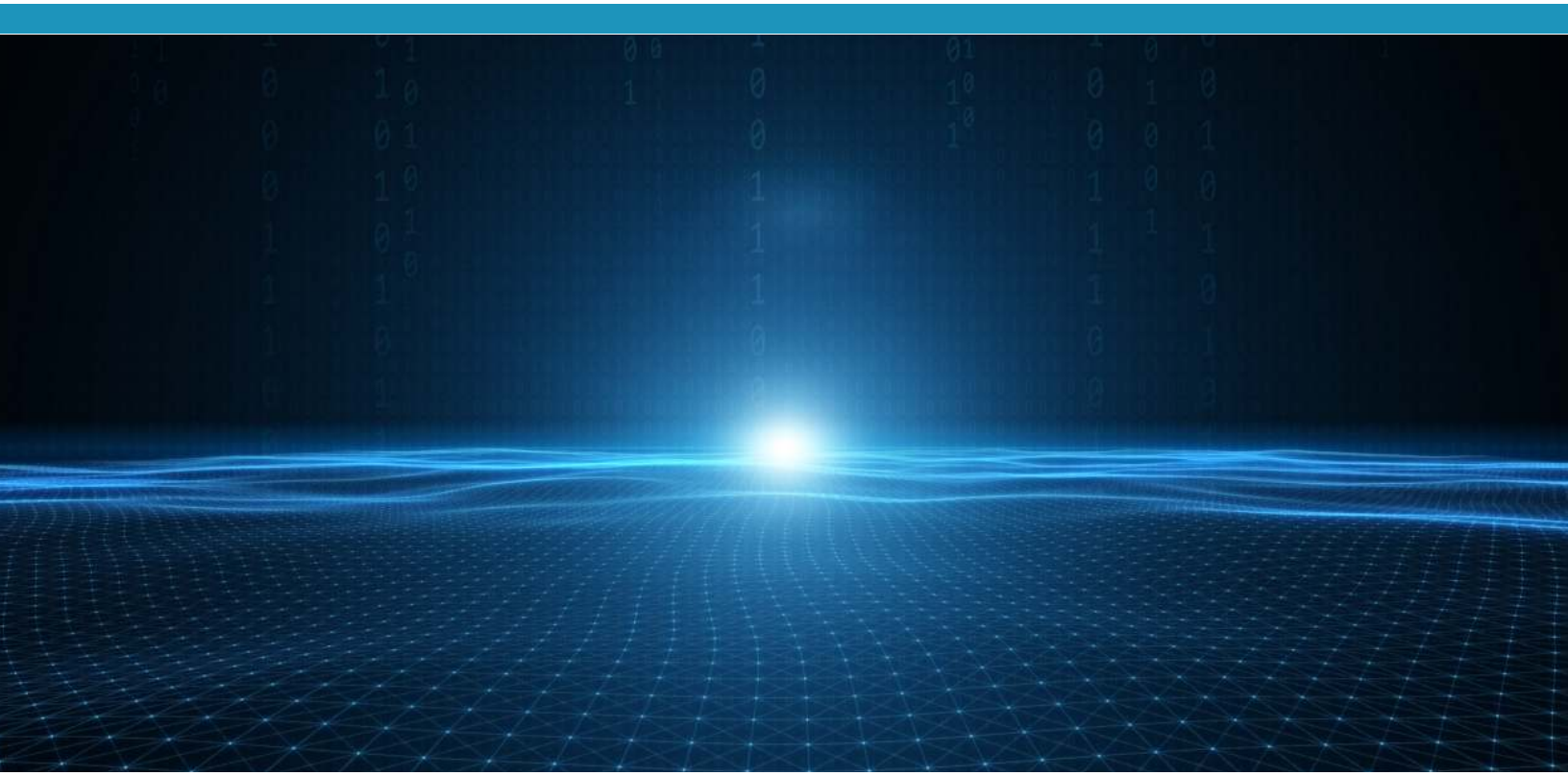


Quality Control (QC) Analysis with PEAKS 11

Yiming Xiao, PhD, Application Scientist
Qing Zhang, MSc, Software Systems Manager
Kyle Hoffman, PhD, Applications Manager

Bioinformatics Solutions Inc., Waterloo, Canada

Abstract:

PEAKS 11 supports QC analysis to monitor attributes of LC-MS/MS data that fail to meet a specified standard. In this study, we utilized published label-free quantification (LFQ) datasets to effectively showcase the application and functionality of the QC module within PEAKS 11.

Introduction:

LC-MS-based protein/peptide quantification has garnered considerable attention in the fields of physiopathology and pharmaceutical research. In particular, clinical studies rely on a substantial number of samples to ensure adequate statistical power, often necessitating over 50 samples per group due to the significant interindividual variation [1]. For cohorts with such large sample sizes, the implementation of a rigorous Quality Control (QC) step becomes indispensable for unobstructed and reliable statistical analyses. Only through this rigorous approach can experimental conclusions be confidently validated.

Addressing this need, PEAKS 11 introduces a specialized QC module that enhances protein/peptide identification and quantification results. The module offers sophisticated QC analysis, ensuring the presentation of interactive and user-friendly outputs tailored to the specific requirements of the user's research.

Case study: benchmarking data set analysis with PEAK 11:

A published data set [2] was used as an example to demonstrate the capability of label free quantification (LFQ) and QC analysis in PEAKS 11.

Study aims and background:

LC-MS-based protein/peptide quantification has been a subject of continuous interest in the fields of physiopathology and pharmaceutical studies. Label-free quantification (LFQ) allows for different MS acquisition strategies, namely, data-dependent acquisition (DDA) and data-independent acquisition (DIA). While DDA represents the more traditional approach, it is susceptible to high missing values and under-sampling. On the other hand, DIA offers increased robustness and can mitigate biases associated with DDA. However, DIA also presents challenges, such as a high false positive rate in complex MS2 spectra.

In this investigation, identical samples were analyzed using high-resolution DDA and DIA (SWATH) methods, and subsequently compared. The quantification of high-resolution DDA (HS-DDA) demonstrated comparable accuracy and precision to DIA, and in certain cases, even outperformed DIA for proteins characterized by low abundance and small fold-changes.

For the purpose of this application note, only DDA data was utilized to demonstrate the capabilities of the QC module in PEAKS 11. This choice was made to emphasize the software's comprehensive QC analysis and its adeptness in presenting results effectively.

Experimental design:

Five groups of samples were prepared with three different proteomes (human, E.coli and yeast) and run with five technical replicates (n=25). The human protein amount proportion was 60% across all samples. The portion of E. coli and yeast protein amounts were as follows:

A: 5%/35%, B: 7.5%/32.5%, C: 10%/30%, D: 15%/25%, E: 20%/20%.

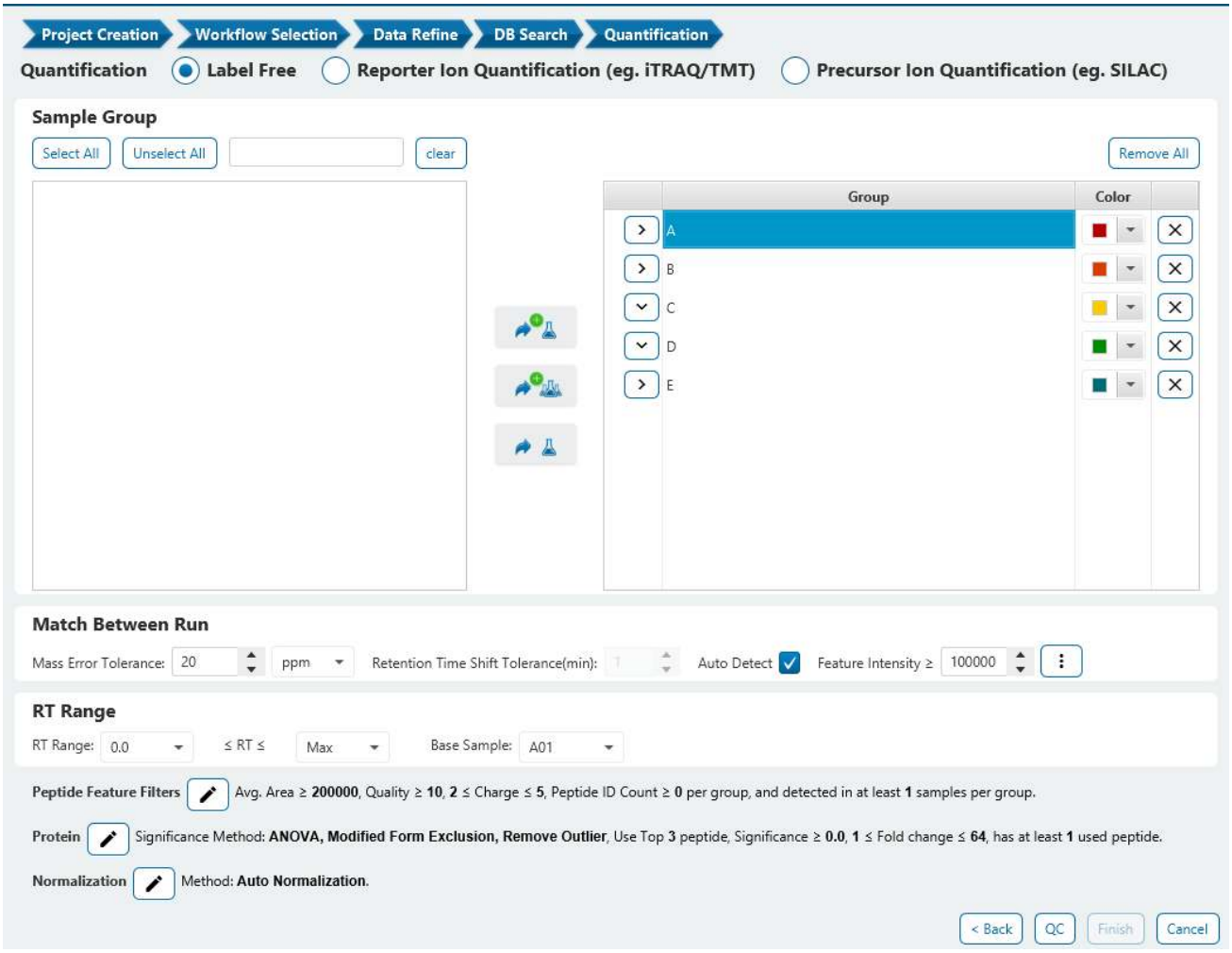
DDA data was acquired with MS1 240000 resolution and MS2 15000 resolution. The gradient was 180 min long. A detail of LC-MS method can be found in [2].

LFQ Data Analysis:

MS data (5 runs *5 samples) was analyzed in PEAKS 11 as LFQ with PEAKS Q module. LFQ was applied with the implementation of match-between-run and total ion current (TIC) normalization techniques. Further specifics on the search parameters and the configuration of the quality control (QC) analysis are presented in Figure 1a (LFQ) and Figure 1b (QC).

While using at least two peptides per protein for quantification, 5157 protein groups are quantifiable. Statistics of filtered results are shown in Figure 2a; the volcano plot is shown in Figure 2b.

1a.



1b.

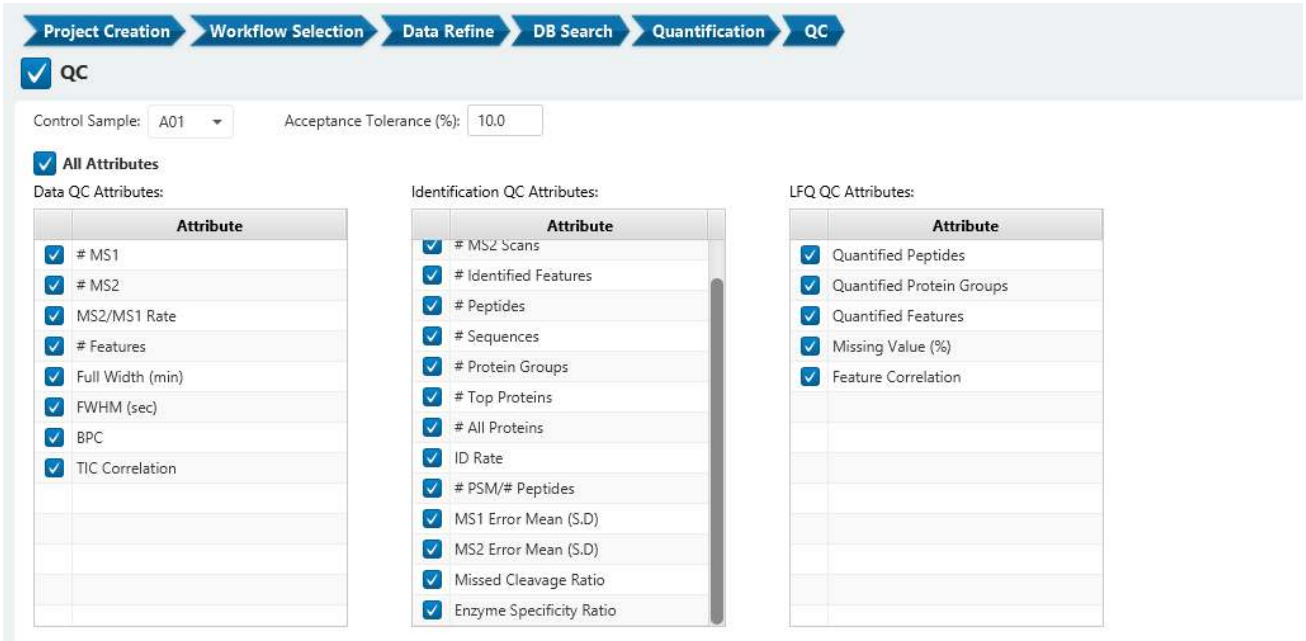


Fig 1. **Establishing protein/peptide LFQ and QC in PEAKS Studio 11** (a) The LFQ setup and parameters in PEAKS Q workflow. Top 3 peptides were selected to calculated protein quantification and auto (TIC) normalization was applied. (b) QC Setup: The QC process was configured by examining all QC attributes. The control sample, A01, was designated for reference, and a tolerance threshold of 10% was set to monitor variations within the QC samples.

QC result of LC-MS data:

In the QC Summary tab, the QC statistics of the LFQ results were categorized into three groups: control sample, passed samples, and failed samples. Samples were classified as failed if any of their attributes fell outside the tolerance level, while samples with all attributes within +/- 10% of the control sample (A01) were considered as passed samples. A concise overview of the QC findings is presented in Figure 3(a). Notably, among the 29 samples in the dataset, only 10 samples exhibited all QC attributes falling within the 10% tolerance range.

Figure 3(b) displays the distribution of the failed attributes in this dataset, with "TIC correlation" and "Feature correlation" being the two most prevalent. These two attributes collectively imply potential concerns regarding the dataset's reproducibility. Further elaboration and discussion on these specific QC attributes will be provided in subsequent sections.

2a.

Features	1625959
Features with ID	917380
Feature vectors with ID	65386
Protein groups	5157

2b.

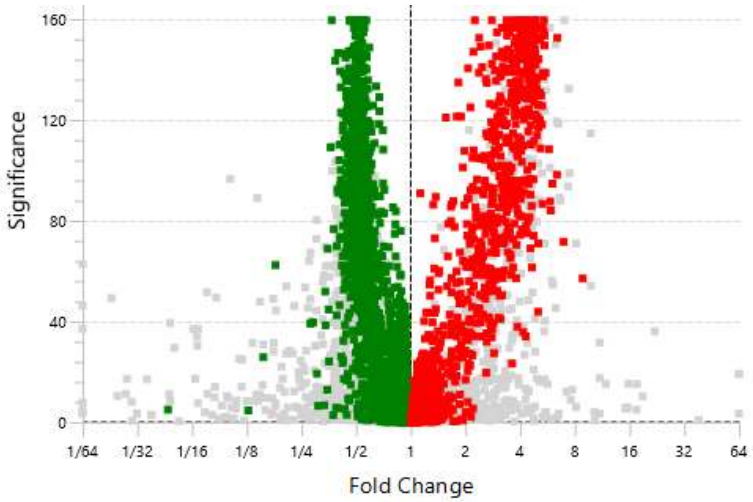


Fig 2. **LFQ result.** (a) statistics of quantification result (b) Volcano plot illustrating differentially expressed proteins. Proteins with downregulation are indicated in green, upregulated proteins in red, and proteins that do not pass the filter criteria (i.e., containing only one quantifiable peptide) are depicted in grey.

3a.

Selected Samples Count	24
Passed Samples Count	10
Ratio of Passed Samples	41.67%
Failed Samples Count	14
Ratio of Failed Samples	58.33%
Total Failed Attributes	6

3b.

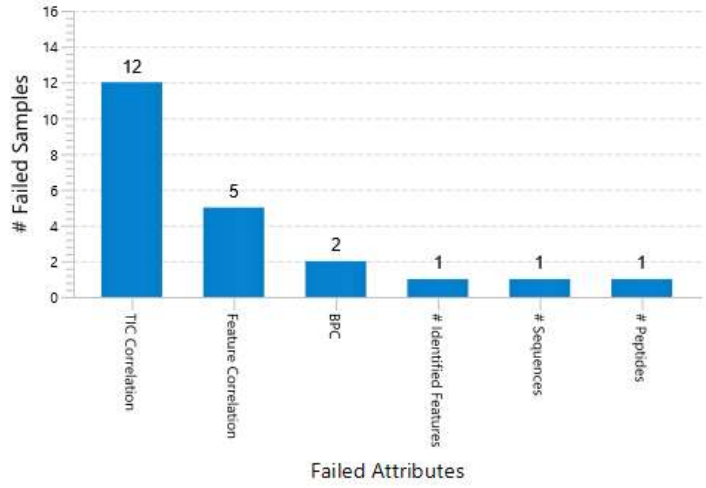


Fig 3. **QC summary of all samples.** (a) statistics of QC results (b) The distribution of failed attributes.

Mass error in QC:

The mass error emerges as a common factor contributing to imperfections in MS datasets, especially when dealing with large cohorts of samples. Environmental temperature fluctuations, MS calibrations, and power surges are potential factors responsible for the occurrence of MS errors. Thus, it is imperative to initially investigate whether the QC attribute failures can be attributed to MS errors. The PEAKS QC module offers multiple avenues for researchers to assess MS errors. The distribution of MS errors was visualized across the categories of control, passed, and failed samples (Figure 4). The violin chart analysis revealed no significant differences between passed and failed samples in both MS1 and MS2 levels. This observation suggests that mass error may not be the primary cause for the samples failing to meet the QC criteria.

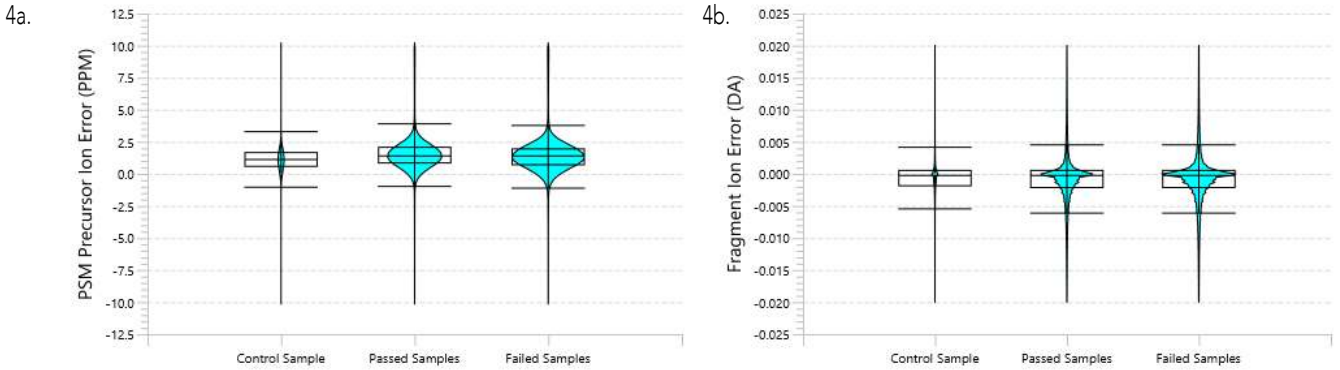


Fig 4. **Sample error distribution.** (a) MS1 Mass Error distribution (b) MS2 Mass Error distribution.

In addition to the collective comparison between passed and failed samples, individual samples' mass errors can also be independently compared with the control sample. Figure 5 demonstrates this comparison, where the sample with the highest MS1 error (D03) was selected for evaluation. It is evident from the analysis that sample D03 remains well within the acceptable mass error range of +5/-5 ppm when compared to the control sample. This further reinforces the notion that the mass error is not a major contributing factor to the failure of samples to meet the QC criteria in the dataset.

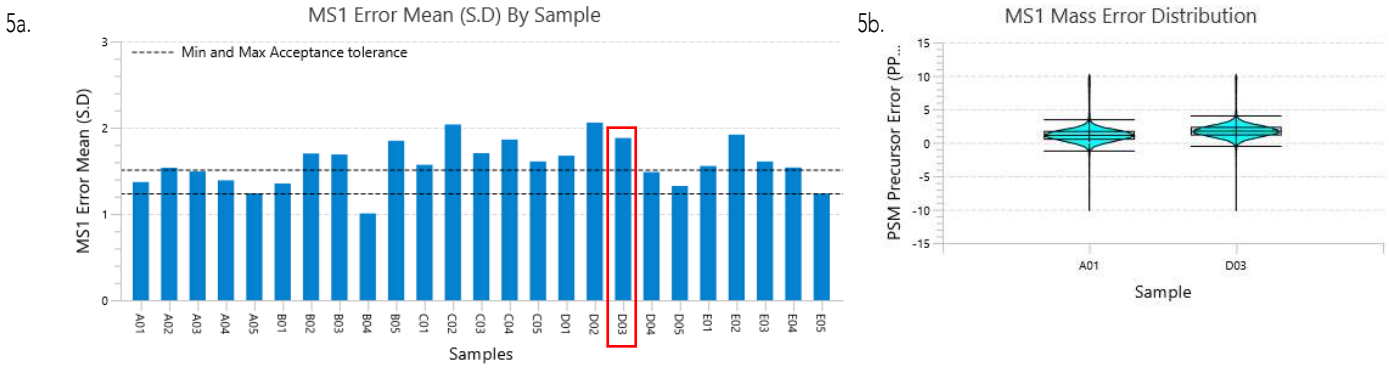


Fig 5. **Individual Sample error distribution.** (a) Bar chart of MS1 error mean of all samples. D03, marked by red rectangles, has the highest MS1 error mean (b) MS1 error distribution comparison between control sample A01 and D03.

TIC Correlation:

Figure 3(b) highlights the most frequently failed QC attribute as TIC correlation. This attribute serves as a measure of similarity between the selected sample and the control sample. In a set of technical replicates, the TIC chromatograms should exhibit high similarity, leading to relatively high TIC correlation values. However, in this study, the TIC correlation is not ideal, even within the same technical replicate datasets (Figure 6).

Figure 7 presents the TIC chromatograms of two samples from technical replicate group A, overlaid with the control sample. In Figure 7(a), the TIC chromatograms of the two samples are nearly identical, indicating high similarity. However, in Figure 7(b), there is a small but discernible difference between the two chromatograms, suggesting reduced similarity.

The evaluation of chromatogram similarity provides a quick and effective method to assess sample reproducibility. In PEAKS 11 QC module, multiple tools are offered to check the identification and quantification reproducibility across various sample aspects. Figure 8 illustrates Venn diagrams depicting the peptide and protein identifications between samples A03 & A01 and A05 & A01. Even between technical replicates (A01 and A03) with highly similar TIC chromatograms, the exact reproducibility is slightly below 80% (common identified proteins/peptides divided by the total): 78% for proteins and 61% for peptides. This observation highlights the stochastic nature of DDA experiments where the precursor selection is semi-random.

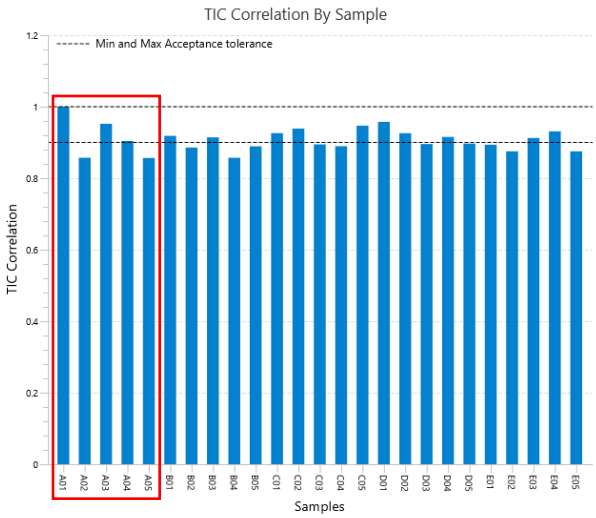


Fig 6. TIC correlation of each sample comparing to control sample A01. The technical replication of group A was marked with red rectangle.

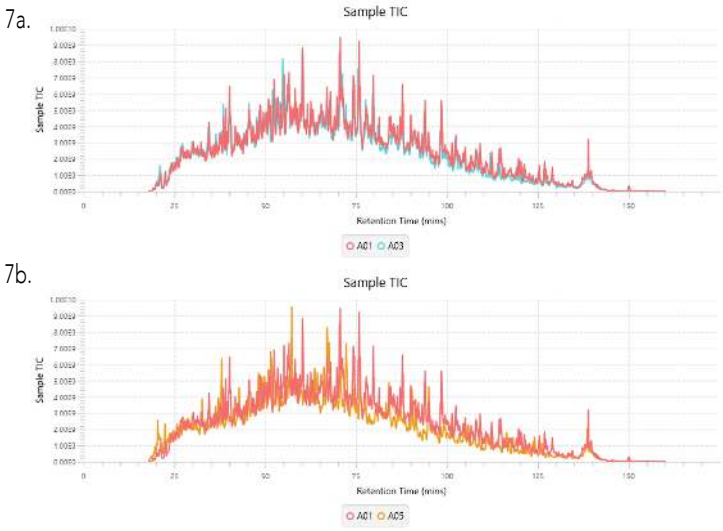


Fig 7. TIC chromatogram overlay between control sample and selected sample (a) A03 vs A01, TIC correlation 0.95 (b) A05 vs A01, TIC correlation 0.86.

The identification reproducibility is somewhat lower between A01 and A05, with 76% for proteins and 54% for peptides. This decrease in reproducibility is understandable since the chromatograms are less identical between sample A01 and A05.

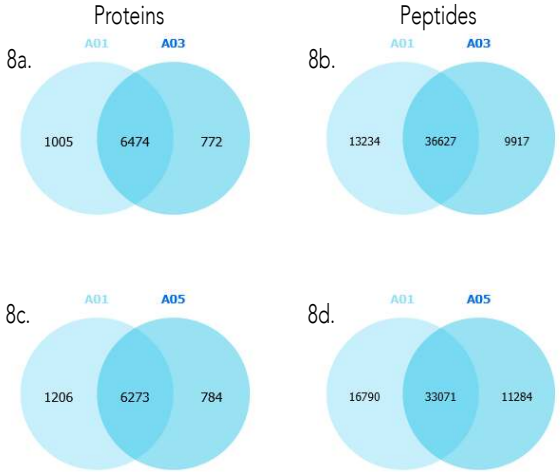


Fig 8. **Venn diagram of sample identification.** (a)(b), number of unique or common identified proteins or peptides in A05 and A01. (c)(d), number of unique or common identified proteins or peptides in A05 and A01.

The evaluation of chromatogram similarity provides a quick and effective method to assess sample reproducibility. In PEAKS 11 QC module, multiple tools are offered to check the identification and quantification reproducibility across various sample aspects. Figure 8 illustrates Venn diagrams depicting the peptide and protein identifications between samples A03 & A01 and A05 & A01. Even between technical replicates (A01 and A03) with highly similar TIC chromatograms, the exact reproducibility is slightly below 80% (common identified proteins/peptides divided by the total): 78% for proteins and 61% for peptides. This observation highlights the stochastic nature of DDA experiments where the precursor selection is semi-random.

In DDA experiments, it is common to observe suboptimal identification reproducibility, even in technical replicates, due to semi-random precursor selection. However, in LFQ experiments, where MS1 ions are used for quantification and ID-transfer is applied, the missing value issues are significantly reduced, and quantification reproducibility should be higher between technical replicates.

Figure 9 illustrates the quantification correlation across three different levels: proteins, peptides, and features. Both A03 and A05 demonstrate excellent linear correlation, exceeding 0.95, when compared to the control sample A01. The quantification correlation in A05 is only marginally lower, approximately 0.01, compared to A03. This finding implies that despite sample A05 not passing the QC attribute check in TIC correlation, its LFQ reproducibility remains largely unaffected.

Notably, the quantification correlations within technical replicates A exhibit high values above 0.95 (data not shown). This observation suggests that our current TIC correlation tolerance level might be overly stringent.

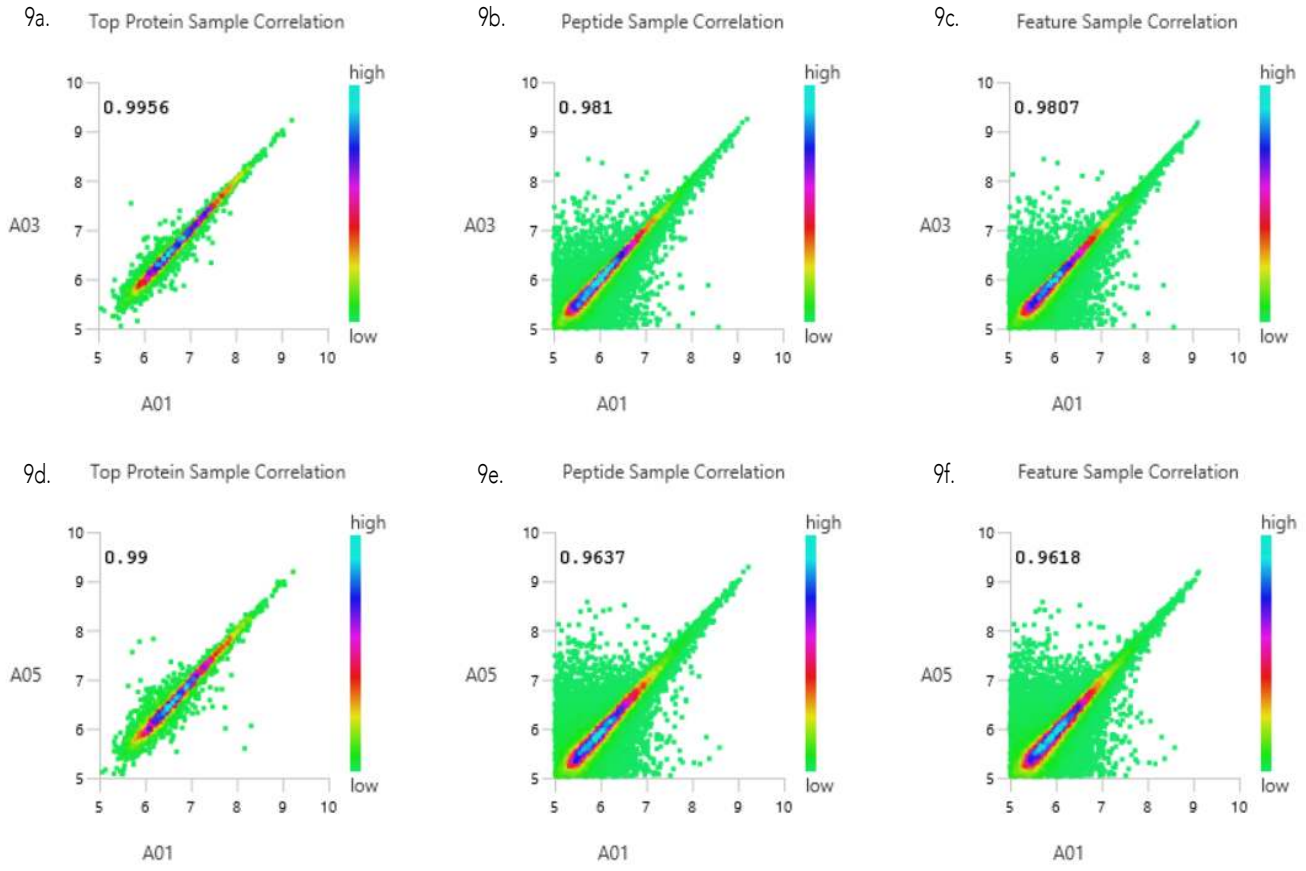


Fig 9. Pearson correlation of quantification results. a-c, proteins, peptides and features correlation between A03 and A01; d-f, proteins, peptides and features correlation between A05 and A01

LC Conditions:

While the overall sample reproducibility remains largely unaffected, it remains of interest to investigate the factors contributing to the observed decrease in TIC chromatogram correlation. One potential reason for the changes in TIC chromatogram correlation could be variations in the LC conditions.

Fortunately, PEAKS 11 QC module offers multiple tools to assess the column performance, gradient suitability, and other relevant factors. In Figure 10(a), it is demonstrated that the full width at half maximum (FWHM) across the entire sample set remains consistent, indicating that the column and LC conditions throughout the run exhibit no significant shifts. Furthermore, the comparison of MS2 count over time and LFQ retention time (RT) alignment between A01 and A05 also suggests that no major LC condition changes occur even between two samples exhibiting TIC chromatogram differences.

These findings indicate that factors other than major LC condition changes may be contributing to the observed variations in TIC chromatogram correlation. Further investigation and analysis are required to identify and address these factors effectively.

MS Sensitvity:

The observed variations in TIC chromatogram correlation may also be influenced by changes in MS sensitivity during sample acquisition. Accumulation of neutral particles on the front end of the MS can lead to decreased instrument sensitivity, affecting the TIC chromatogram and resulting in low TIC correlation.

Figure 11 provides intriguing insights, revealing a trend with the base peak chromatogram (BPC) and the number of identified peptides across different samples within a technical replicate group. The decreasing trend in both BPC and peptide identification numbers suggests a potential decline in MS sensitivity.

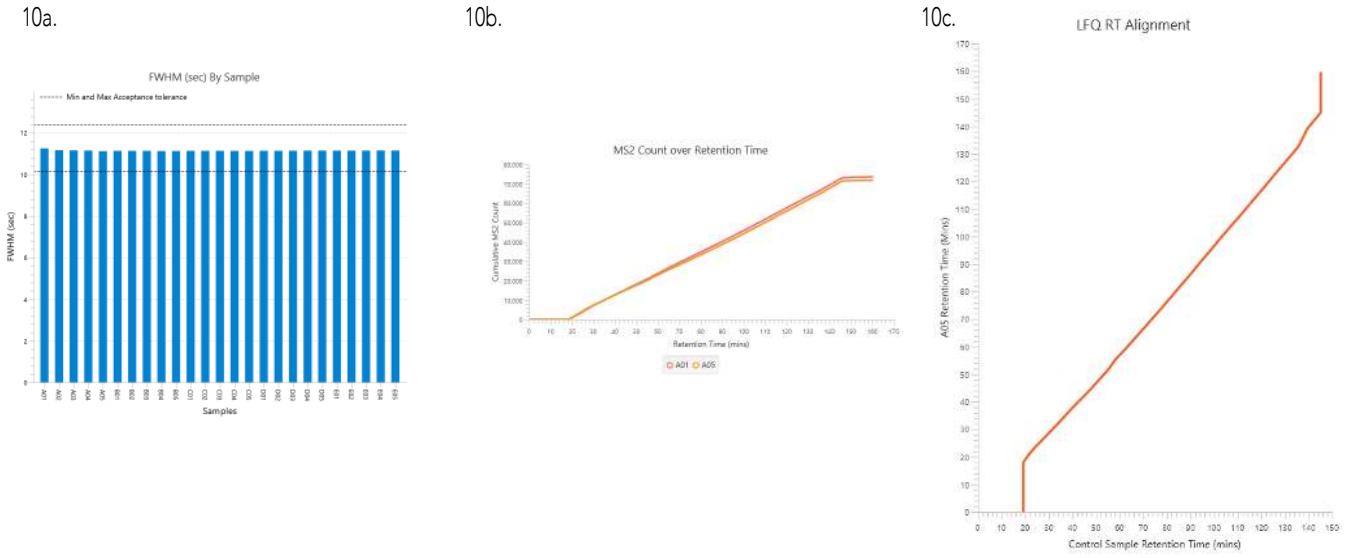


Fig 10. **Quality control of LC condition changes.** (a) FWHM of all samples (b) MS2 over retention time comparison between A01 and A05 (c) LFQ RT alignment between A01 and A05.

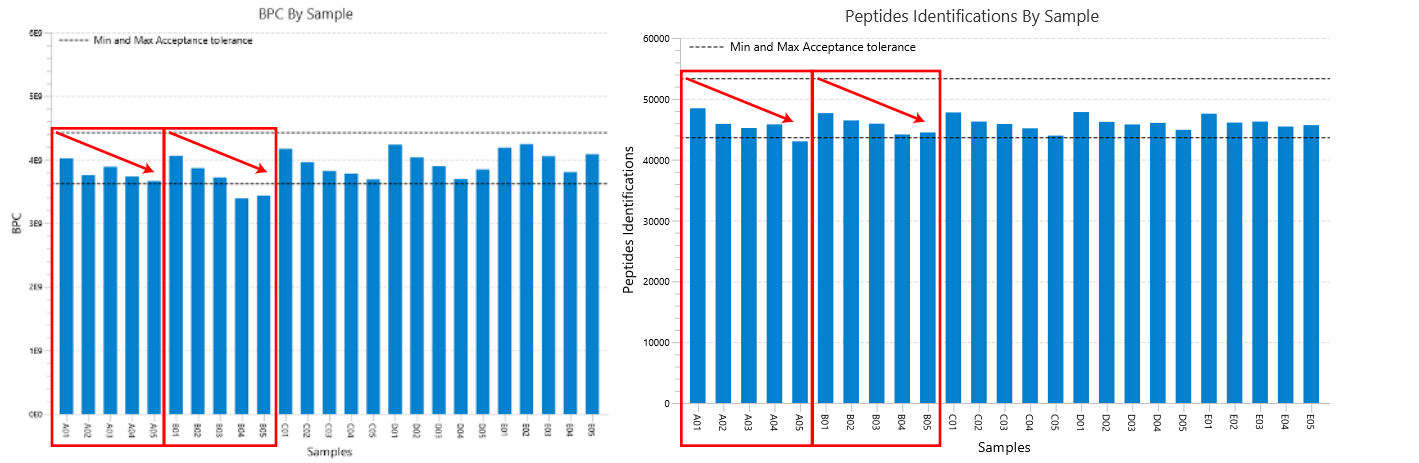


Fig 11. .(a) FWHM of all samples (b) Peptide identification numbers of all sample. The rectangle and the arrow show the decreasing trend in technical replicates.

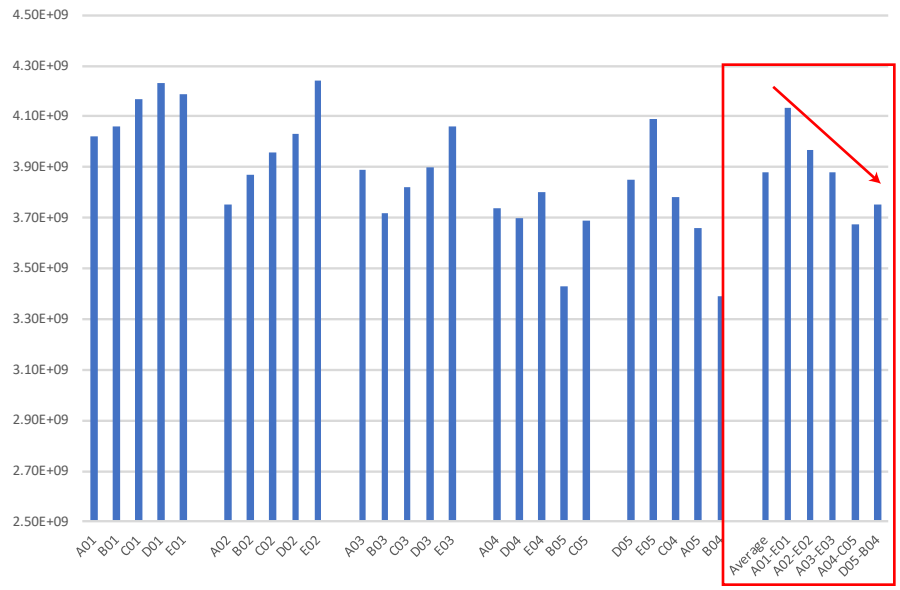
To explore the underlying reasons for this trend, we regrouped the samples based on their acquisition time (Figure 12(b)). By plotting the BPC of each sample and the average BPC of different time groups (Figure 12(a)), a clear trend emerges. The average BPC of time groups indicates a gradual decrease in MS sensitivity over time, explaining the observed differences in TIC chromatograms within the technical replicate groups.

Feature Correlation:

The feature correlation emerges as another crucial QC attribute that falls short of meeting the threshold, and this phenomenon is exclusively observed in samples belonging to group E (Figure 13(a)). This observation indicates the existence of inherent differences between the technical groups, particularly related to the variation in sample amounts between yeast and E. coli (as indicated in the experimental design).

To gain further insights into these population differences, Figure 13(b) reveals three distinct populations representing higher E. coli protein levels and lower yeast protein levels in sample E05. This observation is consistent with the protein density ratio depicted in Figure 13(c). The observed discrepancy in protein levels between yeast and E. coli in sample E05 aligns with the variations observed in the feature correlation, potentially contributing to the failure of this QC attribute in the group E samples..

12a.

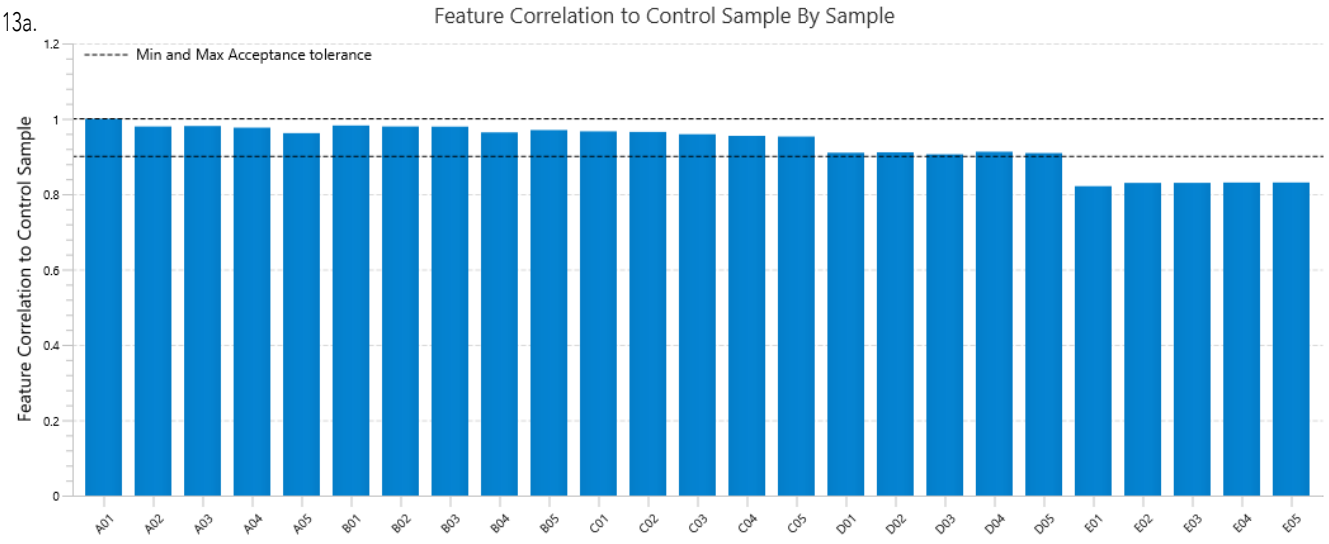


12b.

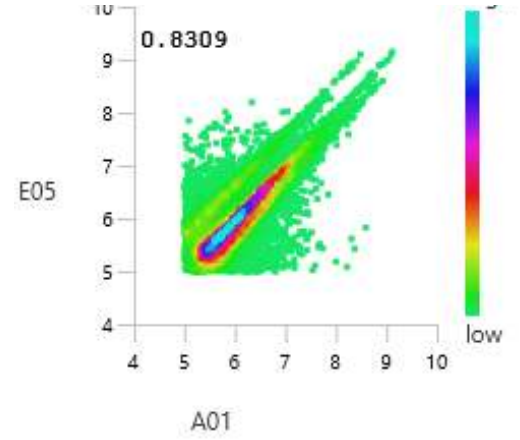
Sample	Acquisition time
A01	4/26/2018 5:38:48 PM
B01	4/26/2018 11:40:38 PM
C01	4/27/2018 2:41:29 AM
D01	4/27/2018 5:42:20 AM
E01	4/27/2018 8:41:11 AM
A02	4/27/2018 2:44:55 PM
B02	4/27/2018 8:46:43 PM
C02	4/27/2018 11:47:33 PM
D02	4/28/2018 2:48:23 AM
E02	4/28/2018 5:49:14 AM
A03	4/28/2018 11:50:58 AM
B03	4/28/2018 5:52:40 PM
C03	4/28/2018 8:53:30 PM
D03	4/28/2018 11:54:18 PM
E03	4/29/2018 2:55:10 AM
A04	4/29/2018 8:56:56 AM
D04	4/29/2018 9:00:16 PM
E04	4/30/2018 12:01:06 AM
B05	4/30/2018 12:04:36 PM
C05	4/30/2018 3:05:28 PM
D05	4/30/2018 6:06:24 PM
E05	4/30/2018 9:07:14 PM
C04	5/2/2018 12:14:58 AM
A05	5/3/2018 3:36:39 PM
B04	5/3/2018 9:28:25 PM

Fig 12. (a) BPC of samples grouped by acquisition time. The red rectangle highlights the average BPC of different groups. (b) the list of data acquisition time of all samples.

13a.



13b.



12c.

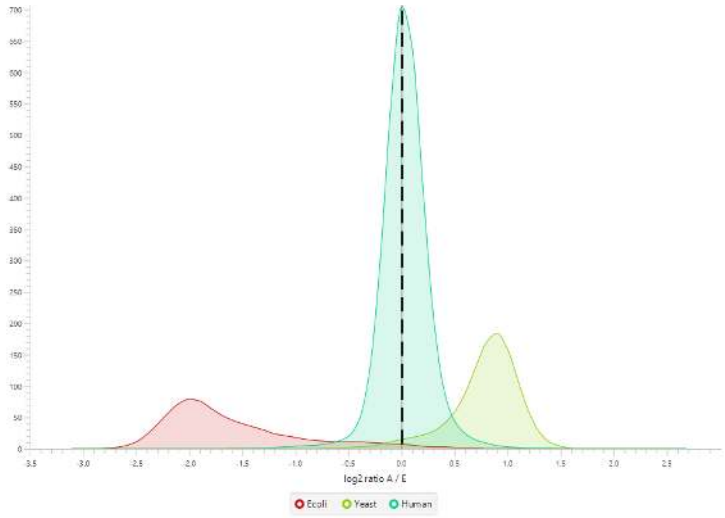


Fig 13. (a) All samples' feature correlation to control sample (b) The feature correlation distribution between E05 and A01 (c) Protein density ratio between group A and group E

Conclusion:

This study employed PEAKS 11 to conduct Label-Free Quantification (LFQ) and Quality Control (QC) analyses on a diverse dataset. The investigation focused on changes in TIC chromatogram and revealed that it was primarily caused by a drop in MS sensitivity. While the sample reproducibility was not significantly affected in this study, it highlights the importance of consistently monitoring machine status, including factors like MS error, MS sensitivity, and LC conditions. The QC analysis in PEAKS 11 proved invaluable in identifying and understanding potential issues influencing sample reproducibility and data quality. By addressing these factors, we can draw robust and meaningful conclusions from our dataset, furthering our understanding of the underlying biological processes under investigation.

References:
1. Overmyer, K. A. et al. Large-scale multi-omic analysis of COVID-19 severity. Cell Syst. 12, 23–40 (2021).
2. Wang, X. et al. Ultra-High-Resolution IonStar Strategy Enhancing Accuracy and Precision of MS1-Based Proteomics and an Extensive Comparison with State-of-the-Art SWATH-MS in Large-Cohort Quantification. Analytical Chemistry. 11, 4884-4893 (2021).

Bioinformatics Solutions, Inc.

140 Columbia St, Suite 202
Waterloo, Ontario N2L 3K8
Canada

Tel: (519) 885-8288
Fax: (519) 885-9075

sales@bioinfor.com
www.bioinfor.com