# Development of Bioinformatics Platform for Analyzing MS-based Protein Identification and Quantification

**Jin-Young Cho**

**Integrated OMICS for Biomedical Science**
**World Class University**
**The Graduate School**
**Yonsei University**

# Development of Bioinformatics Platform for Analyzing MS-based Protein Identification and Quantification

A Dissertation
Submitted to the Department of
Integrated OMICS for Biomedical Science of
World Class University and the Graduate School of Yonsei University
In partial fulfillment of the
Requirements for the degree of
Doctor of Philosophy

Jin-Young Cho

December 2016

**This certifies that the dissertation of Jin-Young Cho is approved.**

---

**Thesis Supervisor: Dr. Young-Ki Paik**

---

**Thesis Committee: Dr. Ho Jeong Kwon**

---

**Thesis Committee: Dr. Jong-Bok Yoon**

---

**Thesis Committee: Dr. Jaewhan Song**

---

**Thesis Committee: Dr. Jong Shin Yoo**

**The Graduate School**
**Yonsei University**

**December 2016**

# 감사의 글

2008년 석사과정 학생으로 대학원을 진학하여 실험실 생활을 시작한 지가 엊그제 같은데, 벌써 박사과정을 마친다고 생각하니 감회가 새롭습니다. 학부 과정을 마치고 바로 석사 과정을 시작할 당시에는 부족함이 많았지만 그저 새로운 환경에서 내가 진짜 하고 싶은 공부를 할 수 있다는 생각에 들떠 있었습니다. 지금 돌이켜보면 그간 배운 것도 많고 성장도 하였지만, 그만큼 많은 한계를 경험한 시간이었습니다. 막연히 가지고 있던 꿈과 기대는 현실 앞에서 수도 없이 깨졌고, 그만큼 새로운 목표와 꿈을 다시 그려 왔습니다. 무수한 시도와 노력 끝에 걸음마를 시작한 아기처럼, 이 논문은 이제 막 박사 과정을 마치고 학문적 걸음마를 시작한 제 첫 징표이기에 더 소중하게 여겨집니다.

이 자리를 빌어 제가 걸음마를 시작할 수 있도록 지도 교수로써 격려를 아끼지 않으신 백융기 교수님께 깊이 감사 드립니다. 교수님의 지원과 조언, 그리고 인내 어린 가르침 덕에 여기까지 올 수 있었다고 생각합니다. 처음 석사 과정으로 대학원에 합격하여 지도 교수 및 전공에 대한 상담을 위해 학교를 찾았을 때, 따뜻이 맞아주시고 백융기 교수님을 소개해 주신 이원태 교수님, 학생을 가르침에 있어 항상 열정과 인자함으로 대해 주신 구현숙 교수님, 석사 과정 당시 졸업에 많은 도움을 주신 김우택 교수님, 석사 과정 당시 학위논문 심사를 맡아 주시고, 이후에도 WCU 융합오믹스의생명과학과의 주임 교수로써 많은 조언과 도움을 주신 조진원 교수님, 부족함이 많았던 석사 논문과 박사 논문 심사에 적극적으로 도움을 주시고 조언을 아끼지 않으신 권호정 교수님, 졸업 시험과 학위논문 심사에 적극적으로 도움을 주신 윤종복 교수님과 송재환 교수님, 그리고 제 학위논문 심사를 위해 그 먼 길을 마다치 않으시고 달려와 주신 유종신 박사님께 깊은 감사를 드립니다.

함께 한 시간은 불과 몇 달이라는 짧은 기간이었지만, 처음 들어와 아무것도 모르는 제게 기본적인 교육과 연구 방향을 잡는데 많은 조언을 주신 민석이 형, 그리고 학위 과정을 수료하는 동안 때로는 친한 형처럼, 때로는 학문적 멘토로서 도움과 조언을 아끼지 않으신 슬기 형은 제게 좋은 롤모델이자 버팀목이 되어 주셨습니다. 함께 석사 과정 동기로 시작하여 동고동락하며 항상 동료이자 인생

선배로서 든든한 버팀목이 되어준 주완이 형께도 감사의 마음을 전합니다. 지금은 각각 바쁜 일정을 보내고 계실, 선임 연구원으로서의 훌륭한 리더십을 보여주신 조상연 박사님과 이은영 박사님, 실험과 학문에 대한 끊임없는 열정을 보여주시고 제게 무엇과도 바꿀 수 없는 실험 데이터와 조언을 아끼지 않으신 형주 형, 묵묵히 본인의 바쁜 연구 스케줄을 소화하시면서도 항상 친절하게 후배들을 챙기시고 자상하게 지도해 주시는 근이 형, 친근한 인상으로 좋은 대화 상대와 조언을 해 주신 안성이 형, 분석 시료를 전해 드리러 인하대병원에 방문할 때마다 실험을 비롯하여 많은 조언과 도움을 주신 광렬이 형, 한 가정의 아내로서, 어머니로서, 연구원으로서 1인 3역을 거뜬히 수행해 내시면서도 항상 웃음으로 센터를 이끌어 주시는 슈퍼우먼 은영, 혜영 누나, 온갖 궂은 일과 과제 일을 거뜬히 소화해 내면서도 성실하고 꼼꼼하게 소임을 다하는 민정이, 석사 과정을 무사히 마치고 졸업하여 이제 한 가정의 어머니로서 열심히 살아가고 있는 선희, 같은 연구실 동료로 동고동락하며 가정을 이루고, 부모로서, 학생으로서 맡은 바 역할에 충실하고 성실한 현정이와 헌이, 형주 형의 뒤를 이어 MS 팀장으로 소임을 다하며 실험과 관련하여 나에게 많은 도움을 준 종선이, 함께 동고동락 하였지만, 지금은 진로를 바꾸어 다른 길을 간 한호, 이제 막 학위 과정을 시작하며 막내로서 온갖 잡무와 힘든 일이 많을 텐데도 불평 한번 안하고 항상 의욕적이고 야무지게 해내는 채연이와 윤진이, 온갖 행정업무로 바쁘겠지만 꼼꼼히 업무를 잘 수행하시고 무엇보다 우리들의 연구비와 각종 금전적 지원에 차질이 없도록 애써 주시는 김민서 선생님과 최선희 선생님 등 모든 YPRC 식구들께 감사 드립니다. 늘 친근한 미소로 맞아 주시는 희경이 형, 스케줄상 자주 찾아 뵙지 못함에도 반갑게 맞아주시던 주효진 박사님과 함정훈, 이정의, 김선희 선배, 그리고 나래, 혜림이, 새람이, 준영이 등 여러 AGPL 식구들께도 감사 드립니다.

특히 제게 물심양면으로 지원을 아끼지 않으시고, 힘들고 방황하는 제게 무한한 사랑과 인내, 그리고 기회를 제공해 주신 어머니, 아버지께 감사 드리고, 또 감사 드립니다. 집안일과 두 아들의 양육을 도맡아 하시면서도 각종 봉사활동과 사회활동, 그리고 화가로서 제 2의 인생을 의욕적으로 살고 계신 어머니, 퇴직 후에도 일손을 놓지 않으시며 기다림과 격려로 용기를 북돋아 주신 아버지가 없었더라면 지금의 저는 없었을 것입니다. 항상 근면하고 우리 집안의 분위기 메이커인 동생 진호에게 형으로써 큰 힘이 된 적이 없는 것 같아 항상 미안하고 고맙다는 말을 전하고 싶습니다. 각자의 환경에서 맡은 역할을 충실히 하며 힘들 때마다 위로와 힘이

되어주는 내 친구 남흥이와 승훈이, 훈민이에게도 고맙다는 말을 전합니다.

어린시절 그렇게 저를 귀여워해 주셨지만 고생만 하고 가신 것 같아 안타까운 할머니, 군복무 중 운명을 달리 하셔서 끝까지 임종을 지켜드리지 못한 할아버지께 죄송하고 감사 드립니다. 두 분이 주셨던 사랑은 항상 간직하도록 하겠습니다.

끝으로 이렇게 좋은 사람들과 인연을 맺어 주시고, 기회와 환경을 부여해 주신 하나님께 감사 드립니다. 비록 지금 제가 떼어놓은 걸음은 미약하고 불안하지만, 그간 여러분이 보내 주신 격려와 사랑을 양식으로 학문적, 신앙적, 사회적으로 한 사람의 몫을 다할 수 있는 인재가 되도록 노력하겠습니다. 모든 분들께 진심으로 감사의 말씀 드립니다.

2016년 12월

조진영 드림

# Contents

## SUBJECT II:
## Epsilon-Q: An Automated Analyzer Interface for Mass Spectral Library Search and Label-Free Quantification

# List of Tables

## SUBJECT I:

## A Combination of Multiple Spectral Libraries Improves the Current Search Methods Used to Identify Missing Proteins in the Chromosome-centric Human Proteome Project



## SUBJECT II:

## Epsilon-Q: an automated analyzer interface for mass spectral library search and label-free quantification

# List of Figures

## Introduction



## SUBJECT I:

## A Combination of Multiple Spectral Libraries Improves the Current Search Methods Used to Identify Missing Proteins in the Chromosome-centric Human Proteome Project

## SUBJECT II:

## Epsilon-Q: an automated analyzer interface for mass spectral library search and label-free quantification

# Abbreviations

| | |
|---|---|
| **C-HPP** | The Chromosome-centric Human Proteome Project |
| **CID** | Collision-induced Dissociation |
| **DB** | Database |
| **FDR** | False Discovery Rate |
| **FWHM** | Full Width at Half Maximum Peak Height |
| **HPLC** | High Performance Liquid Chromatography |
| **iRefSPL** | Integrated Reference Peptide Spectral Library |
| **ISB** | The Institute for Systems Biology |
| **m/z** | Mass-to-Charge Ratio |
| **MAD** | Median Absolute Deviation |
| **MS** | Mass Spectrometry |
| **MS/MS** | Tandem Mass Spectrometry |
| **NIST** | The National Institute of Standards and Technology |
| **PSM** | Peptide-Spectrum Match |
| **PXD** | ProteomeXchange Database |
| **refSPL** | Reference Peptide Spectral Library |
| **simSPL** | Simulated Peptide Spectral Library |
| **UPS** | The Universal Proteomics Standard |
| **XIC** | Extracted Ion Chromatogram |

# I. Abstract

## Development of Bioinformatics Platform for Analyzing MS-based Protein Identification and Quantification

**Jin-Young Cho**

**Department of Integrated OMICS for Biomedical Science of World Class University**

**The Graduate School**

**Yonsei University**

Approximately 2.9 billion long base-pair human reference genome sequences are known to encode some 20,000 representative proteins. However, 3,000 proteins, i.e., about 15% of all proteins, have no or very weak proteomic evidence and still missing, termed missing protein. Missing proteins may be present in rare samples at very low abundance or with only temporary expression, causing some problems in their detection for protein profiling. In particular, some technical limitations cause those missing proteins remain unassigned. For example, current mass spectrometry (MS) techniques have detection limits and high error rates for complex biological samples. Insufficient proteome coverage of a reference sequence database (DB) and a spectral library also major issues. Thus, the development of a better search strategy that results in greater sensitivity and more accurate in search of missing proteins is necessary. To this end, we used a new

strategy, which combines a reference spectral library searching and a simulated spectral library (simSPL) searching to identify missing proteins. We built the human iRefSPL, which contains the original human reference spectral library and additional peptide sequence-spectrum match entries from other species. We also built the human simSPL, which contains simulated spectra of 173,907 human tryptic peptides by MassAnalyzer (version 2.3.1).

To prove the enhanced analytical performance of the combination of human iRefSPL and simSPL method, called "Combo-Spec Search method", for the identification of missing proteins, we attempted to re-analyze the placental tissue dataset (PXD000754). Each experiment data was analyzed by PeptideProphet, and the results were combined by iProphet. For the quality control, we applied class-specific false-discovery rate (FDR) filtering method. All results were filtered at less than 1% FDR in peptide and protein level. The quality controlled results were cross-checked with the neXtProt DB (2014-09-19 release). The two spectral libraries, iRefSPL and simSPL were designed to have no overlapped proteome coverage. They showed complementary in spectral library searching and significantly increased the number of matches. From this trial, 12 missing proteins were newly identified, which passed the criterion—Least two of 7 or more length amino acid peptides or one of 9 or more lengths amino acid peptide with one or more unique sequence. Thus, the use of the iRefSPL and simSPL combination can be helped to identify peptides that had not been detected by conventional sequence DB searches with improved sensitivity and low error rate.

We developed a new analytical software, called Epsilon-Q. This software is designed to support Combo-Spec Search and label-free quantification method. Epsilon-Q supports standard MS data format and connects with SpectraST to match spectrum-to-

spectrum. Epsilon-Q automatically performs three operations: raw MS data indexing, multiple spectral library searching and calculating sum of precursor ion peak intensities for user input datasets. By using the multi-threading function, Epsilon-Q can performs multiple spectral library searching and parsing the results. With user friendly graphical interface, Epsilon-Q has shown a good performance to identify and quantify proteins. Especially, for low abundance proteins in biological samples, Epsilon-Q has outperformed other sequence DB search engines. Thus, we anticipate that Epsilon-Q software helps users to get improved detectability in identifying proteins and to perform comparative analysis of biological samples.

Keyword: Bottom-up Proteomic Approach, Chromosome-centric Human Proteome Project, Combo-Spec Search method, Epsilon-Q, Label-free Quantification, Mass Spectrometry, Missing Protein, Protein Identification, Proteomics, Sequence Database Search, Spectral Library

# II. Introduction

A bottom-up proteomic approach is commonly used to identify proteins by mass spectrometry (MS) analysis coupled with high-pressure liquid chromatography (HPLC) (Aebersold, *et al.*, 2003; Chait, 2006). The proteins are extracted from the samples and digested by a protease(s) (e.g., trypsin) to produce a peptide mixture. The mixture is subsequently injected into the reverse-phase HPLC. While the peptides passed through the column, it is separated by its physicochemical properties (i.e. hydrophobicity, charge, and molecular size). The molecular ions of each peptide are then introduced into the mass spectrometer. The ions are fragmented, frequently by collision-induced dissociation (CID), and their mass-to-charge ratio (m/z) and intensity are recorded in subsequent MS/MS spectra. The MS/MS spectra are used as a query to identify the peptides and subsequently the proteins in the sample (see figure 1).

## Two MS data Analytical Methods for Protein Identification

Sequence database (DB) searching (Steen, *et al.*, 2004; Zhang, *et al.*, 2014) is the most widely used method for MS-based proteomics (Craig, *et al.*, 2004; Eng, *et al.*, 1994; Geer, *et al.*, 2004; Liu, *et al.*, 2004; Perkins, *et al.*, 1999; Tabb, *et al.*, 2007). Sequence-to-spectrum matching in the method is performed by automated sequence DB search tools such as SEQUEST (Eng, et al., 1994), MASCOT (Perkins, et al., 1999), X!TANDEM (Fenyo, *et al.*, 2003), MyriMatch (Tabb, et al., 2007) and MS-GF+ (Kim, *et al.*, 2014) (see figure 2A). However, in this approach, only m/z values are used to sequence-spectrum

matching and any other spectral information, such as residue-specific effects in cleavage and variable fragment mass peak intensities, are ignored. It may cause low sensitivity and potential errors in the handling of low-quality experimental spectra, especially those contaminated by any polymer or other noise peaks (Yen, *et al.*, 2011) (see red box of figure 2A).

Spectral libraries have been used for the MS-based identification of small molecules since the 1980s (Lam, *et al.*, 2011; Stein, *et al.*, 1994). Spectral library searching takes all of the spectral features into accounts, such as peak intensities, the natural loss of fragments, and various unknown fragments that are specific to certain peptides (see figure 2B). Thus, spectral library searching shows greater sensitivity and better matching of results than sequence DB searching (Craig, *et al.*, 2006; Lam, *et al.*, 2007). Yates *et al.* (Yates, *et al.*, 1998) suggested that this approach could be used for the identification of peptides and proteins. Spectral library searching was recently reported to outperform sequence DB searching (Hu, *et al.*, 2013; Lam, *et al.*, 2008; Zhang, *et al.*, 2011). Spectral library search algorithms and software, such as SpectraST (2007)(Lam, et al., 2007), X!Hunter (2006)(Craig, et al., 2006), and BiblioSpec (2006)(Frewen, *et al.*, 2006), were released at around the same time and are now widely used in this approach. The National Institute of Standards and Technology (NIST) now provides reference spectral libraries for humans and eight other species. The PeptideAtlas, developed by the Institute for Systems Biology (ISB), provides almost 61 million human peptide spectra and various spectral libraries of individual human organisms (e.g., the brain, heart, kidney, liver, and plasma)(Desiere, *et al.*, 2005).

## Limitations of Spectral Library Searching

To build a spectral library, the accumulation of data is essentially, which is depending on high-quality tandem MS spectra with high-scored peptide sequence assignment by stringent quality control criteria. It promises reliability of spectral library, but this is also one of reasons why the spectral library has low proteome coverage and slowly increasing data accumulation rate than sequence DB (Hu, *et al.*, 2011). Usually, peptide spectral library has lower proteome coverage than protein sequence DB (see figure 3). Several strategies have been proposed to expand the proteome coverage of the reference spectral library by including the predicted spectra of unobserved peptides (Yen, et al., 2011; Yen, *et al.*, 2009). For example, it has been suggested that the fragmentation patterns of a peptide in MS can be predicted by its sequence and physicochemical properties (Zhang, 2004; 2005). The CID spectra of similar peptides show extremely similar intensity patterns, which implies that the MS spectra of a peptide can be predicted by the neighbor-based approach based on its sequence (Ji, *et al.*, 2013). Information-driven semi-empirical spectra of the reference spectral library were also demonstrated to be useful for the detection of novel phosphorylated peptides (Hu, et al., 2011; Suni, *et al.*, 2015a).

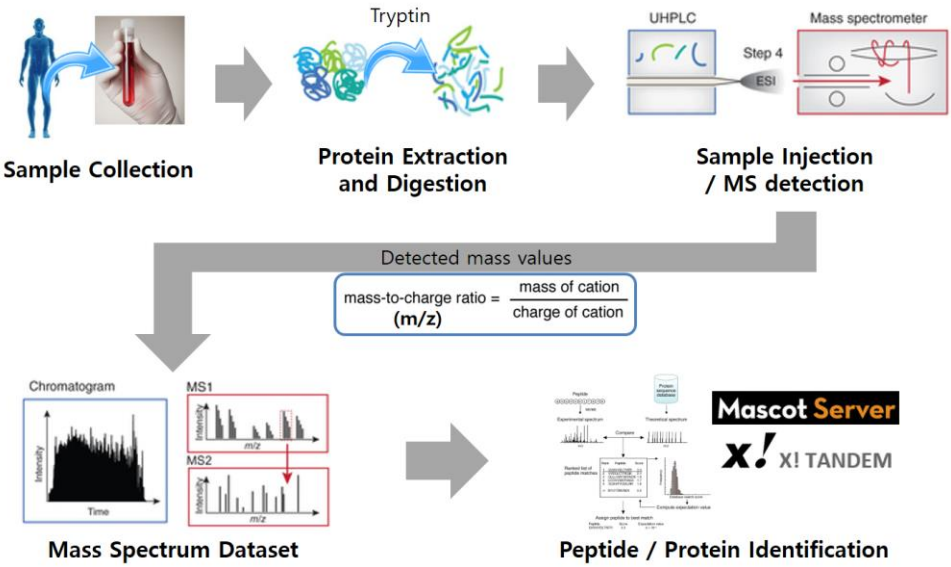



**Figure 1.** Mass spectrometry-based bottom-up proteomic approach. To detect proteins by this approach, each protein is digested by a protease(s) (e.g., trypsin) to produce a peptide mixture. The peptides are then injected into mass spectrometer and detected for their m/z value. Using the m/z values and analytical software, we can identify protein sequences in target sample.

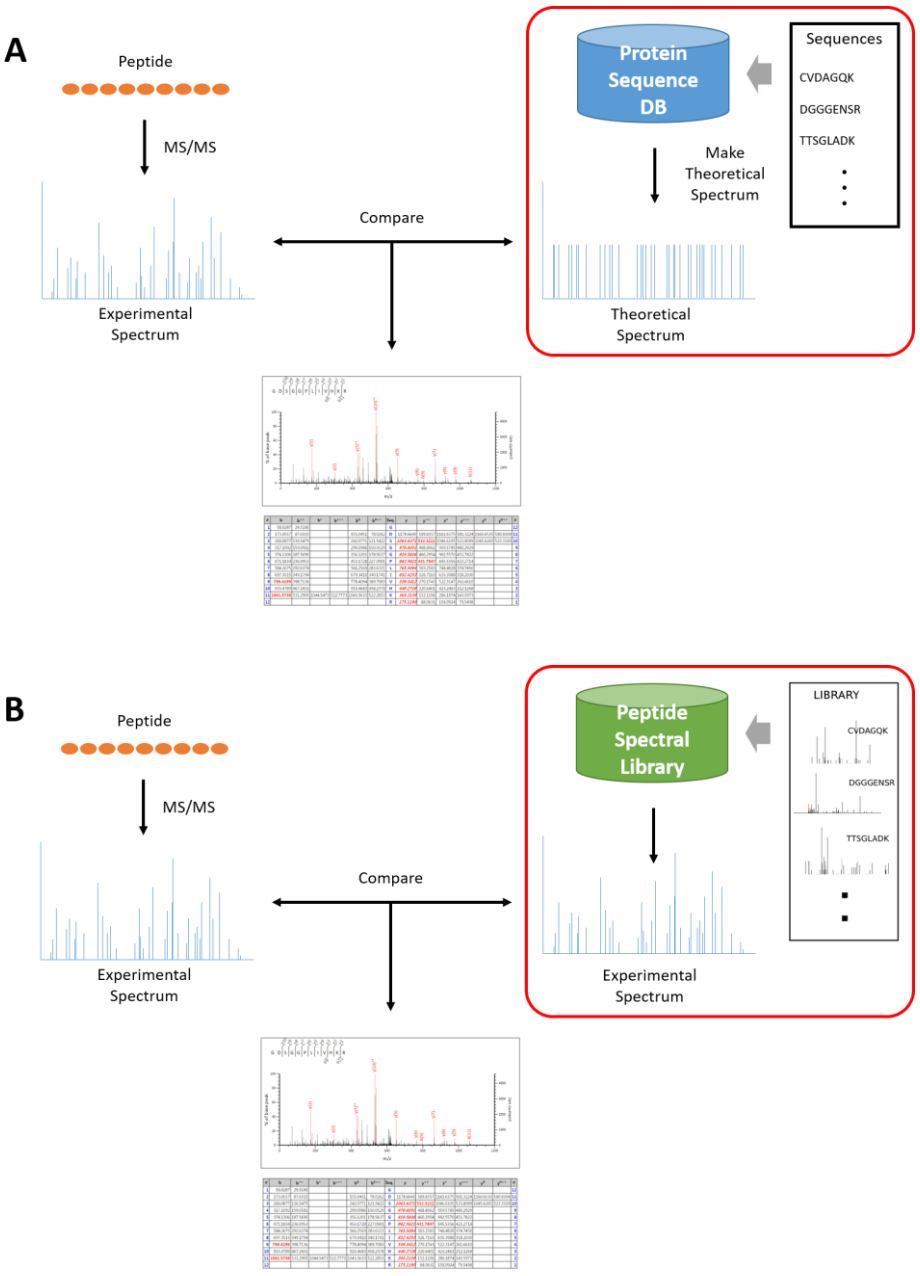



**Figure 2.** Two mass spectrometry data analytical methods for protein identification. (A) Sequence database searching workflow. (B) Peptide spectral library searching workflow.

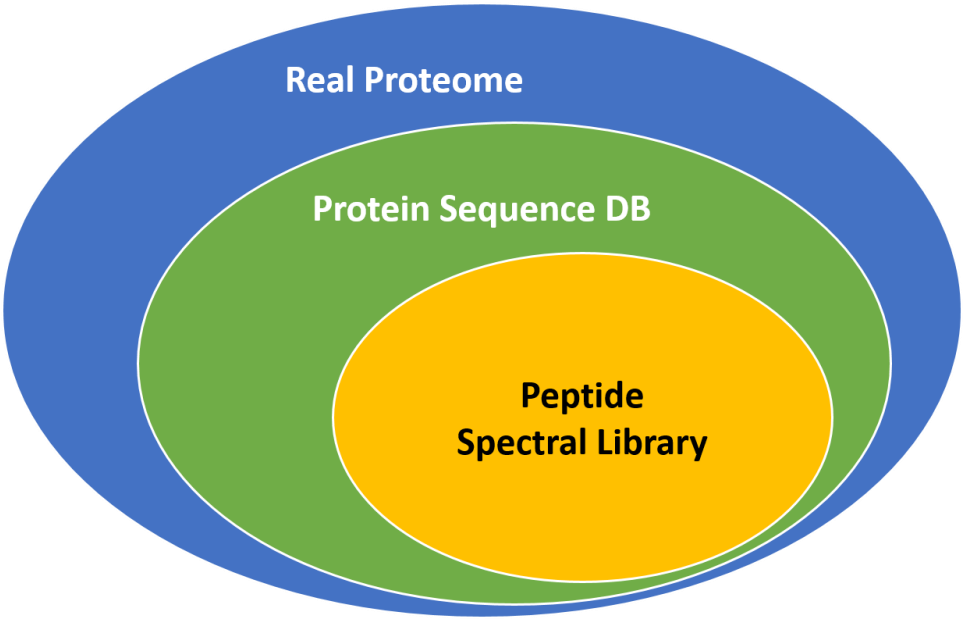



**Figure 3.** Comparison of proteome coverages between sequence database and peptide spectral library. Blue circle means real proteome coverage in biological sample (100%) whereas green circle represents proteome coverage of protein sequence database. Yellow circle indicates proteome coverage of peptide spectral library.

## SUBJECT I

# A Combination of Multiple Spectral Libraries Improves the Current Search Methods Used to Identify Missing Proteins in the Chromosome-centric Human Proteome Project

# 1. Introduction

Approximately 2.9 billion long base-pair human reference genome sequences are now known to encode some 20,000 representative proteins (Maher, 2012). By inference, many proteins are not only directly encoded by a genome sequence but also diversified by the additional processing such as the post-transcriptional and post-translational modification. The direct analysis of cell and tissue protein expression is, therefore, necessary to collect a list of parts (Dhingra, *et al.*, 2005; Gygi, *et al.*, 1999). The Chromosome-centric Human Proteome Project (C-HPP) consortium was founded to map and annotate all of the proteins that are encoded by genes on each of the chromosomes found in humans (Paik, *et al.*, 2012a; Paik, *et al.*, 2012b). A total of 25 C-HPP working groups from 20 nations integrate proteomics data into a genomic framework and annotate human proteins using a range of unique and often rare clinical samples. All of the currently available techniques are used to improve our understanding of complex human biological systems and disease states. However, despite the efforts of the teams, about 3,000 proteins still have no clear proteomic evidence (supported by mass spectrometry [MS] or antibody detection). These proteins have been colloquially termed “missing proteins” (Lane, *et al.*, 2014; Paik, et al., 2012a; Paik, et al., 2012b).

A bottom-up proteomic approach is commonly used to identify proteins by MS analysis coupled with high-pressure liquid chromatography (HPLC) (Aebersold, et al., 2003; Chait, 2006). The proteins are extracted from the samples and digested by a protease(s) (e.g., trypsin) to produce a peptide mixture. The mixture is subsequently injected into the reverse-phase HPLC. While the peptides passed through the column, it

is separated by its physicochemical properties (i.e. hydrophobicity, charge, and molecular size). The molecular ions of each peptide are then introduced into the mass spectrometer. The ions are fragmented, frequently by collision-induced dissociation (CID), and their mass-to-charge ratio (m/z) and intensity are recorded in subsequent MS/MS spectra. The MS/MS spectra are used as a query to identify the peptides and subsequently the proteins in the sample.

Sequence database (DB) searching (Steen, et al., 2004; Zhang, et al., 2014) is the most widely used method for MS-based proteomics (Craig, et al., 2004; Eng, et al., 1994; Geer, et al., 2004; Liu, et al., 2004; Perkins, et al., 1999; Tabb, et al., 2007). Sequence-to-spectrum matching in the method is performed by automated sequence DB search tools such as SEQUEST (Eng, et al., 1994), MASCOT (Perkins, et al., 1999), X!TANDEM (Fenyo, et al., 2003), MyriMatch (Tabb, et al., 2007) and MS-GF+ (Kim, et al., 2014). However, in this approach, only m/z values are used to sequence-spectrum matching and any other spectral information, such as residue-specific effects in cleavage and variable fragment mass peak intensities, are ignored. It may cause low sensitivity and potential errors in the handling of low-quality experimental spectra, especially those contaminated by any polymer or other noise peaks (Yen, et al., 2011).

Spectral libraries have been used for the MS-based identification of small molecules since the 1980s (Lam, et al., 2011; Stein, et al., 1994). Spectral library searching takes all of the spectral features into accounts, such as peak intensities, the natural loss of fragments, and various unknown fragments that are specific to certain peptides. Thus, spectral library searching shows greater sensitivity and better matching of results than sequence DB searching (Craig, et al., 2006; Lam, et al., 2007). Yates *et al.* (Yates, et al., 1998) suggested that this approach could be used for the identification of peptides and

proteins. Spectral library searching was recently reported to outperform sequence DB searching (Hu, et al., 2013; Lam, et al., 2008; Zhang, et al., 2011). Spectral library search algorithms and software, such as SpectraST (2007)(Lam, et al., 2007), X!Hunter (2006)(Craig, et al., 2006), and BiblioSpec (2006)(Frewen, et al., 2006), were released at around the same time and are now widely used in this approach. The National Institute of Standards and Technology (NIST) now provides reference spectral libraries for humans and eight other species. The PeptideAtlas, developed by the Institute for Systems Biology (ISB), provides almost 61 million human peptide spectra and various spectral libraries of individual human organisms (e.g., the brain, heart, kidney, liver, and plasma)(Desiere, et al., 2005).

To build a spectral library, the accumulation of data depends on high-quality tandem MS spectra with high-scored peptide sequence assignment by stringent quality control criteria. It promises reliability of spectral library, but this is why the spectral library has low proteome coverage and slowly increasing data accumulation rate than sequence DB (Hu, et al., 2011). Several strategies have been proposed to expand the proteome coverage of the reference spectral library by including the predicted spectra of unobserved peptides (Yen, et al., 2011; Yen, et al., 2009). For example, it has been suggested that the fragmentation patterns of a peptide in MS can be predicted by its sequence and physicochemical properties (Zhang, 2004; 2005). The CID spectra of similar peptides show extremely similar intensity patterns, which implies that the MS spectra of a peptide can be predicted by the neighbor-based approach based on its sequence (Ji, et al., 2013). Information-driven semi-empirical spectra of the reference spectral library were also demonstrated to be useful for the detection of novel phosphorylated peptides (Hu, et al., 2011; Suni, et al., 2015a).

In this study, we describe a new strategy, which uses a combination of multiple spectral libraries (e.g., a reference spectral library and a simSPL) for spectrum-spectrum matching to identify the proteins of interest in cell or tissues. We demonstrate that, compared with conventional sequence DB searching methods, the method can provide improved sensitivity and lower error rate to identify missing proteins by extended proteome coverage.

# 2. Materials and Methods

## 2.1. Datasets

The datasets, which used in this study, were obtained from the ProteomeXchange database (PXD). First, we obtained dataset files that generated by 48 purified human recombinant proteins mixture (UPS, Sigma-Aldrich, St. Louse, MO. USA) in spiked into the biological sample (published by Ahrné *et al.*, PXD000331)(Ahrne, *et al.*, 2013). We used the dataset, called the UPS dataset, to evaluate the performance and effectiveness of our approach. Second, we used the MS dataset obtained from human placental tissue that was previously analyzed by Lee *et al.* (PXD000754)(Lee, *et al.*, 2013). This dataset was generated using various protein enrichment techniques (ThermoFisher LTQ Orbitrap) and MS for the comprehensive proteomic analysis of human placental tissue. We used this dataset to re-analyze and evaluate our new method for the search for novel peptides that are possibly derived from missing proteins. The more detailed metadata of the datasets is in table 1.

## 2.2. Integration of human reference spectral library (iRefSPL)

The reference spectral libraries were obtained from PeptideAtlas (ISB) and the NIST public library repository. We selected the libraries that contained the only CID-fragmented ion spectra, as listed in table 2. All obtained human reference spectral libraries were combined as a consensus spectral library (human refSPL). Proteome coverage of the original human refSPL was expanded by extracting peptide-spectrum

match (PSM) entries from other species spectral libraries. Because each PSM entries of spectral libraries from PeptideAtlas and NIST has already been validated, we did not put a limit on the maximum sequence length. Thus, the PSM entries from the non-human spectral library were selected by the human tryptic peptide list. The peptide list contains minimum 7 amino acids with a maximum of 2 missed cleavage sites, generated from the SwissProt human protein sequence DB (2015-04). All impure spectra were removed or marked by SpectraST software (Version 5.0, Build 201408281759-6544:6594M by Henry Lam). All of the selected PSM entries were added to human refSPL to build a human iRefSPL.

## 2.3. Generation of simulated spectral library (simSPL)

We obtained 41,061 protein sequences from neXtProt (2014-09-19). We compiled a tryptic peptide list of the proteins, as mentioned above, with a length of 7 to 35 amino acids, and a maximum of 2 missed cleavage sites. Total 2,227,896 sequences were selected for the simulation of their MS/MS spectrum. MassAnalyzer (version 2.3.1) was used to simulate MS/MS spectrum of the selected peptides. The simulation parameters were: Orbitrap instrument profile; CID fragmentation mode; isolation width, 2.5; resolution, 800; collision energy (V), 35; and activation time, 30 ms. We considered two charge states: +2 and +3 precursors. We added two types of modification into the simulated spectra: carbamidomethylation at cysteine residues for fixed modifications and oxidation at methionine residues for variable modifications. The predicted spectra were converted to the *.splib format by SpectraST (Hu, et al., 2013). All PSM entries which already included in iRefSPL were removed. The simulated spectral library was called the “human

simSPL".

## 2.4. Protein identification and data analysis

All MS data files were converted into "mgf" and "mzXML" formats by msconvert (Build date: June 17, 2013). Three protein sequence DB search engines were used for sequence DB searching: Mascot Server (version 2.2.07, Matrixscience), X!Tandem (2013.06.15.1 – LabKey, Insilicos, ISB), and Comet (version 2014.02 rev. 2, University of Washington). The sequence DB search parameters were: trypsin for protein digestion, carbamidomethylation at cysteine residues (+57 Da) for fixed modifications, oxidation at methionine (+16 Da) for variable modifications, a maximum of two missed cleavages, 5 ppm MS tolerance, and 0.6 Da MS/MS tolerance. Two charge states, 2+ and 3+, were considered. To filter the false discovery rate (FDR), reversed protein sequences were included in the target sequence DB using the TOPPAS DecoyDatabase builder (version 1.11.1)(Junker, *et al.*, 2012). SpectraST was used for spectral library build and searching. All results were excluded which had lower F-value than 0.45. To estimate the FDR, we generated an equal-size artificial decoy library and appended it to the target spectral library following the method described by Lam *et al.* (Lam, *et al.*, 2010). Each experiment result was analyzed by PeptideProphet (Keller, *et al.*, 2002) and all the results were combined by iProphet (built in Trans-Proteome Pipeline version 4.8.0 PHILAE, Build 201411201551-6764)(Shteynberg, *et al.*, 2011) with default parameters. We used decoy hits and non-parametric model to pin down the negative frequency. We determined two peptide probability thresholds by class-specific FDR filtering (Nesvizhskii, 2014). Each threshold was determined in separate FDR estimation in two classes (resulted peptide hits

by iRefSPL as a class I and by simSPL as a class II). The FDR of each class was limited less than 1%.

# 3. Results

## 3.1. Construction of the integrated reference spectral library (iRefSPL) which contains peptide spectrum matches from humans and eight non-human species

We designed a method that uses two spectral libraries to expand proteome coverage for spectral library searching and detect additional peptides (Figure 1). To expand the proteome coverage of human reference spectral library, we prepared an integrated reference spectral library, called the iRefSPL. The library was built by combining the original human reference spectral library and PSM entries obtained from the other species spectral libraries. The rationale for this approach was provided by a previous report indicating a close correlation between the peptide fragmentation pattern and the sequence, the state of charge, and modifications (Zhang, 2004; 2005). We expected that the proteome coverage of the spectral library of interest could be expanded by the additional PSM entries and it may not incurring false-positive problem. To estimate the dependence of the fragmentation pattern on the physicochemical properties of the peptide (e.g., sequence, charge state, and modification) through various spectral libraries, we selected common PSM entries from the NIST human reference spectral library and eight other species spectral libraries. A total of 77,056 PSM pairs were collected to compare its similarity through various spectral libraries. The similarity of the PSM pairs was estimated by the dot scoring method (Lam, et al., 2007). Table 3 outlines the distributions of PSM pairs as expressed by their dot scores. Many PSM pairs tend to show close to dot score of 1, suggesting that peptide fragmentation and peak intensity patterns

were highly correlated to their sequence, charge and modification state. Based on the result, we extracted total 51,374 PSM entries from 13 non-human spectral libraries to expand proteome coverage of human refSPL (see table 4). We added the PSM entries, obtained from the 13 non-human species spectral libraries, into the human refSPL to produce human iRefSPL.

To test the effectiveness of added PSM entries, we analyzed placental tissue dataset using both human iRefSPL and human refSPL (called Combo-Spec Search method). Figure 2A shows a prediction model in which the estimated sensitivity and error rate of both the human iRefSPL and the human refSPL. The two results did not differ significantly. By using the human iRefSPL, more peptides were identified, especially in low error rate (<= 0.0005), then human refSPL (Figure 2B). The results suggest that PSM entries that extracted from other spectral libraries can be used to expand proteome coverage of the human refSPL without any false-positive problems.

## 3.2. Comparison of various search methods in sensitivity over error rate and time to processing large MS dataset

We examined the performance of Combo-Spec Search method compared to other conventional approaches in identifying additional peptides with low error rate by using the UPS dataset. Three protein sequence DB search engines (Mascot, X!Tandem, and Comet) and original reference UPS spectral library were used as conventional approaches. The FASTA sequence DB and the reference spectral library of the UPS standard protein mix (UPS refSPL) were obtained from the NIST (released, 2011-05-24). We did not prepared iRefSPL for UPS dataset analysis in this test because the original refSPL from NIST for

UPS dataset analysis has already sufficient proteome coverage (about 85% of the sequences of all of the 48 standard proteins). Thus, we used the refSPL of UPS dataset rather than build additional iRefSPL.

We compared correct matches number through different error rates by refSPL only and three each sequence DB search engines. As we expected that the matches by refSPL only (see top second bar in Figure 3A) shows more increased than the matches that obtained by each single sequence search engine (below three bars in Figure 3A).

The top first bar in Figure 3A shows the effectiveness of the simSPL. The refSPL had 85% of proteome coverage to UPS data, so we build simSPL with the 15% of gaps for complete coverage. We built simSPL which had about 15% of proteome coverage and no overlap with refSPL because the simSPL shows better positive/negative number of sibling peptide distribution in refSPL-simSPL combination than complete proteome coverage version of simSPL (Figure 4).

We suggest that spectral library searching by using the refSPL and simSPL should be performed independently because the libraries has different characteristics. RefSPL has observed spectra and simSPL has simulated spectra. This difference can be occurred different accuracy in spectrum-to-spectrum matching. Usually, refSPL searching shows more accuracy than simSPL searching. So we applied class-specific FDR control before those result integration (figure 5).

In less 1% FDR, we detected 427 different peptides by use of the only refSPL. However, using the combination of simSPL and refSPL, we detected 33 more novel different peptides. The result shows that combination of both refSPL and simSPL (refSPL-simSPL combination method) can more detect peptides in the low error rate than

other conventional methods (refSPL only, single or multiple sequence DB searching). It is known that the use of a combination of multiple search engines would produce highly improved identification rates (Shteynberg, *et al.*, 2013). As known that, the combination of three sequence DB search engines (Multiple DB Search) show significantly increased matches in low error rate (<= 0.0005). To evaluate the sensitivities of both two multiple search strategies (by Multiple DB Search and Combo-Spec Search method), we depicted the relation of sensitivity and error rate. Figure 3B shows that the Combo-Spec Search method shows little more good sensitivity than Multiple DB Search, but it is not significantly different. Both two methods show good sensitivity in various probability thresholds. However, Combo-Spec Search method shows lower error rates than Multiple DB Search in extremely low probability threshold (<= 0.2). This result shows that Combo-Spec Search method has more effective restriction power for errors than Multiple DB Search.

Combo-Spec Search method has shown more reduced time to process MS dataset than other sequence DB search engines (figure 6). The MS dataset (PXD000603) is consisted of 24 raw files and about 41.2GB of size. Because Combo-Spec Search is based on spectrum-to-spectrum matching, it shows less spending time than other sequence DB search engines.

## 3.3. Application of the Combo-Spec Search method to identify missing proteins

To test the performance of the human Combo-Spec Search method in identifying missing proteins, we attempted to re-analyze the human placental tissue dataset

(PXD000754)(Lee, et al., 2013). The dataset was re-analyzed independently by Combo-Spec Search method coupled with SpectraST and the results were combined using iProphet (Figure 7).

The combined results were filtered at an FDR of less than 1% at the protein level. All combined matched results were classified into two groups (matched by human iRefSPL and human simSPL) and separately applied probabilistic threshold (0.8299 for iRefSPL group and 0.9303 for simSPL group) to satisfy less than 1% FDR in peptide level in each group. Figure 8 shows the statistics of the dataset. A total of 4,104 proteins were identified, which was slightly fewer (135) than the previous result of 4,239 proteins (Lee, et al., 2013). It may have been due to the use of CID spectra only in this study various types of the spectrum (CID, higher-energy collisional dissociation, and electron-transfer dissociation) were used in the previous study. The human iRefSPL and simSPL, used in this study, can only support CID type spectra for spectral library searching. By using the multiple sequence DB search engines (Mascot, X!Tandem and Comet), total 3,607 proteins were identified at FDR of less than 1% at the protein level. When the two results that were generated by Multiple DB Search Method and Combo-Spec Search method were compared, the Combo-Spec Search method shows the higher rate of protein identification than the former. When the previous search results (4,239 proteins) were applied to the old version of neXtProt DB (2012-10-07 release), 42 proteins were found to be newly identified missing proteins (Lee, et al., 2013). However, when was applied neXtProt DB (2014-09-19 release) to the Combo-Spec Search Method, 12 proteins were newly found as missing proteins (see table 5 and figure 9). The 12 missing proteins passed our consensus criterion—Least two of 7 or more length peptides or one of 9 or more length peptide with one or more unique sequence. By using the Multiple DB Search Method, there are no

newly identified missing proteins.

The three of all proteins were identified by simSPL. The unique peptides of three proteins were not included in any reference spectral libraries. It is implying that simSPL is complementary to iRefSPL in terms of novel peptide searches. Thus, the use of both iRefSPL and simSPL shows the synergetic effect to identify known and novel peptides from large datasets with high sensitivity and low error rate. It identified peptides that had not been detected by some conventional sequence DB search engines in the previous study. By using the Combo-Spec Search method, we can detect 12 missing proteins from the previously published dataset. It suggests that the method can be useful to re-analyze other previously published data sets and detect additional missing proteins.

# 4. Discussion

Although the rigorous protein search analyzes were carried out on MS data produced under the instrument's optimal performance conditions, it is inevitable that some proteins will remain undetected. It is why we need to develop a better search strategy that provides greater sensitivity and more accurate analysis in the search for missing proteins. Yates *et al.*, suggested that spectral library searching can be a solution to overcome limitations of sequence DB searching (Yates, et al., 1998). According to recent studies, this method outperforms sequence DB searching (Hu, et al., 2013; Lam, et al., 2008; Zhang, et al., 2011). Based on the results, we designed the new method, called "Combo-Spec Search method". This study demonstrates that the application of Combo-Spec Search method to a previously analyzed dataset (Lee, et al., 2013) can provide additional opportunities to identify missing proteins that have never been detected by sequence DB searches. Usually, original reference spectral libraries have insufficient proteome coverage (30-40%) compared to the sequence DB. We suggest that combination of multiple spectral libraries with different proteome coverage could be one solution to overcome the limitation. The improved performance of the Combo-Spec Search method in the identification of missing proteins is due to its expanded proteome coverage. We have shown that Combo-Spec Search method detects more PSMs than other sequence DB search engines and multiple DB search approach. The promising results indicate that it would also be worth reanalyzing already reported datasets deposited in the ProteomeXchange repository in the hope of detecting additional missing proteins. Using the method, we can newly detect 12 missing proteins. There are two olfactory receptors in the 12 missing proteins. It is the exceptional result when considering the sample type

used in this study. We made thorough search again through the currently updated PeptideAtals, but we were not able to find any pieces of evidence for the two olfactory receptors are false-positive matches. However, we do not exclude a possibility of the SNP or any modifications because our newly built spectral libraries (iRefSPL and simSPL) do not contain such rare modification types and SNP. It would be possible to re-examine this issue along with the newly identified 12 missing proteins when the upgraded version of iRefSPL and simSPL that introduces artificial modifications and SNP are available in the future. There are some useful public spectral library and mass spectral data repositories (PeptideAtlas, NIST Peptide Library and GPMdb). The repositories are updated certain intervals (e.g., quarterly or yearly). Using the latest data, we can get more expanded and sophisticated spectral library to be used in the Combo-Spec Search method. Finally, we propose that the Combo-Spec Search method could serve as a common practice in the search for missing proteins and thus could replace the conventional sequence DB search approach.

**Table 1.** Metadata of SUBJECT I study datasets

| | | |
|---|---|---|
| UPS (PXD000331) | Repository | PRIDE |
| | Announce Date | 2014-08-08 |
| | Instrument | LTQ Orbitrap Velos |
| | Contribution | 3 raw files (technical replicate) |
| | Size | Total 25,927 spectra (MS2) |
| | Description | The .raw data submitted to PRIDE correspond to replicate DDA LC-MS/MS analysis of the UPS2 |
| Human Placental tissue profiling (PXD000754) | Repository | PRIDE |
| | Announce Date | 2015-05-26 |
| | Instrument | LTQ Orbitrap |
| | Contribution | 47 raw files (fractions) |
| | Size | Total 266,148 spectra (MS2) |
| | Description | Profilling normal human placantal proteomes using LTQ-OrbiTrap |

**Table 2.** List of the reference spectral libraries used in SUBJECT I study.

| | Library | Fragmentation / Instrument | Build Date | Total number of spectra |
|---|---|---|---|---|
| ISB | Human (Brain) | CID / Iontrap | 2013-08 | 620,813 |
| | Human (Kidney) | CID / Iontrap | 2013-08 | 938,113 |
| | Human (Liver) | CID / Iontrap | 2013-08 | 1,845,053 |
| | Human (Plasma) | CID / Iontrap | 2013-08 | 30,513,825 |
| | Human (Urine) | CID / Iontrap | 2013-08 | 425,579 |
| | Human (Others) | CID / Iontrap | 2013-08 | 29,592,772 |
| | Human (all) | CID / Iontrap | 2013-08 | 61,124,407 |
| | Human (phospho) | CID / Iontrap | 2013-07 | 18,066 |
| | Human (SEMI phospho) | CID / Iontrap | 2013-07 | 35,099 |
| | Mouse | CID / Iontrap | 2013-02 | 4,001,770 |
| | Mouse (phospho) | CID / Iontrap | 2013-07 | 51,420 |
| | *Drosophila* (phospho) | CID / Iontrap | 2013-07 | 16,177 |
| | *C. elegans* | CID / Iontrap | 2013-09 | 1,371,627 |
| | *C. elegans* (phospho) | CID / Iontrap | 2013-07 | 9,225 |
| | Yeast (phospho) | CID / Iontrap | 2013-07 | 18,412 |
| | *Leptospira interrogans* | CID / Iontrap | 2013-08 | 248,430 |
| | Cow | CID / Iontrap | 2011-12 | 196,791 |
| | Honey Bee | CID / Iontrap | 2013-09 | 4,102,541 |
| | *Mtuberculosis* | CID / Iontrap | 2013-07 | 1,134,715 |
| | Pig | CID / Iontrap | 2011-08 | 1,511,129 |
| | Rat | CID / Iontrap | 2013-11 | 2,926,833 |
| NIST | Human | CID / Iontrap | 2014-05 | 340,356 |
| | Mouse | CID / Iontrap | 2013-05 | 149,442 |
| | *Drosophila* | CID / Iontrap | 2012-04 | 78,966 |
| | *C. elegans* | CID / Iontrap | 2011-05 | 67,470 |
| | Yeast | CID / Iontrap | 2012-04 | 50,907 |
| | *E.coli* | CID / Iontrap | 2013-05 | 62,383 |
| | Rat | CID / Iontrap | 2013-05 | 61,707 |
| | Chicken | CID / Iontrap | 2011-05 | 3,125 |
| | Zebrafish | CID / Iontrap | 2015-01 | 28,952 |

**Table 3.** Similarity of common PSM pairs in humans and eight other non-human species* (*Caenorhabditis elegans*, chicken, *Drosophila*, *Escherichia coli*, mouse, rat, yeast, and zebrafish) provided by NIST).

| Similarity (dot score) | C.elegans | Chicken | Drosophila | E.coli | Mouse | Rat | Yeast | Zebrafish |
|---|---|---|---|---|---|---|---|---|
| 1-0.9 | 775 | 6 | 1377 | 19 | 26180 | 11949 | 257 | 909 |
| 0.9-0.8 | 333 | 6 | 734 | 22 | 15203 | 6669 | 130 | 2287 |
| 0.8-0.7 | 92 | 5 | 296 | 4 | 4946 | 2284 | 49 | 839 |
| 0.7-0.6 | 19 | 2 | 58 | 2 | 812 | 495 | 11 | 151 |
| 0.6-0.5 | 1 | 3 | 0 | 0 | 67 | 44 | 1 | 9 |
| 0.5-0.4 | 0 | 2 | 0 | 1 | 4 | 0 | 0 | 2 |
| 0.4-0.3 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 0.3-0.2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0.2-0.1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0.1-0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

**Table 4.** Number of extracted PSM entries from non-human species spectral libraries using the human whole tryptic peptide list.

| | Library | Total # of spectra | Extracted # of spectra |
|---|---|---|---|
| ISB | Mouse | 902,068 | 35,854 |
| | Mouse (phospho) | | |
| | *Drosophila* (phospho) | | |
| | *C. elegans* | | |
| | *C. elegans* (phospho) | | |
| | Yeast (phospho) | | |
| | *Leptospira interrogans* | | |
| | Cow | | |
| | Honey Bee | | |
| | *Mtuberculosis* | | |
| | Pig | | |
| | Rat | | |
| NIST | Mouse | 451,163 | 15,520 |
| | *Drosophila* | | |
| | *C. elegans* | | |
| | Yeast | | |
| | *E.coli* | | |
| | Rat | | |
| | Chicken | | |
| | Zebrafish | | |

**Table 5.** List of identified missing proteins in this study.

| Chr | Protein Accession (Gene name) | Coverage (%) | Total PSMs | Protein Prob. | PE | | |
|---|---|---|---|---|---|---|---|
| | **Peptide Sequence / Charge** | **Length** | **PSMs** | **Peptide Prob.** | **dot** | **F-value** | **Matched library** |
| 1 | Q5VVM6 (CCDC30) | 2.9 | 3 | 0.8953 | 2 | | |
| | DHFLIAC$_{160}$DLLQRENSELETKVLK / 2 | 23 | 3 | 0.8953 | 0.758 | 0.622 | iRefSPL |
| 3 | Q8NGV6 (OR5H6) | 6.8 | 2 | 0.987 | 2 | | |
| | AVSTCGAHLLSVSLYYGPLTFK / 3 | 22 | 2 | 0.987 | 0.898 | 0.783 | iRefSPL |
| 6 | Q8IZF3 (GPR115) | 2 | 88 | 0.9773 | 2 | | |
| | QVNGLVLSVVLPER / 3 | 14 | 88 | 0.9919 | 0.891 | 0.722 | iRefSPL |
| 7 | Q8WXK1 (ASB15) | 2 | 5 | 0.9955 | 2 | | |
| | KGSYDMVSTLIK / 3 | 12 | 5 | 0.9955 | 0.939 | 0.571 | iRefSPL |
| 9 | Q8NE28 (STKLD1) | 3.7 | 3 | 0.8783 | 2 | | |
| | QM$_{147}$VPASITDM$_{147}$LLEGNVASILEVMQK / 3 | 25 | 3 | 0.8783 | 0.713 | 0.607 | iRefSPL |
| 11 | Q6IEU7 (OR5M10) | 3.5 | 11 | 0.9987 | 2 | | |
| | DVILAIQQM$_{147}$IR / 2 | 10 | 11 | 0.9987 | 0.757 | 0.613 | simSPL |
| 13 | O75343 (GUCY1B2) | 2.1 | 2 | 0.9949 | 2 | | |
| | DQEALQAAFLKMK / 3 | 13 | 2 | 0.9949 | 0.908 | 0.698 | iRefSPL |
| 18 | Q9H2F9 (CCDC68) | 5.1 | 5 | 0.9721 | 2 | | |
| | DLQLLEM$_{147}$NKENEVLKIK / 3 | 17 | 5 | 0.9721 | 0.749 | 0.608 | iRefSPL |
| 19 | C9J6K1 (C19orf81) | 7.1 | 8 | 0.9683 | 4 | | |
| | RM$_{147}$LEALGAEPNEEA / 3 | 14 | 8 | 0.9683 | 0.852 | 0.545 | iRefSPL |
| 19 | Q96RP8 (KCNA7) | 3.1 | 3 | 0.9957 | 2 | | |
| | GLQILGQTLRASM$_{147}$R / 3 | 14 | 3 | 0.9957 | 0.816 | 0.623 | simSPL |
| 20 | Q8N687 (DEFB125) | 10.3 | 4 | 0.9243 | 2 | | |
| | NKLSCCISIISHEYTR / 2 | 16 | 4 | 0.9243 | 0.837 | 0.697 | iRefSPL |
| 21 | P57055 (RIPPLY3) | 2.9 | 18 | 0.979 | 2 | | |
| | MEPEAAAGAR / 2 | 10 | 18 | 0.979 | 0.653 | 0.552 | simSPL |

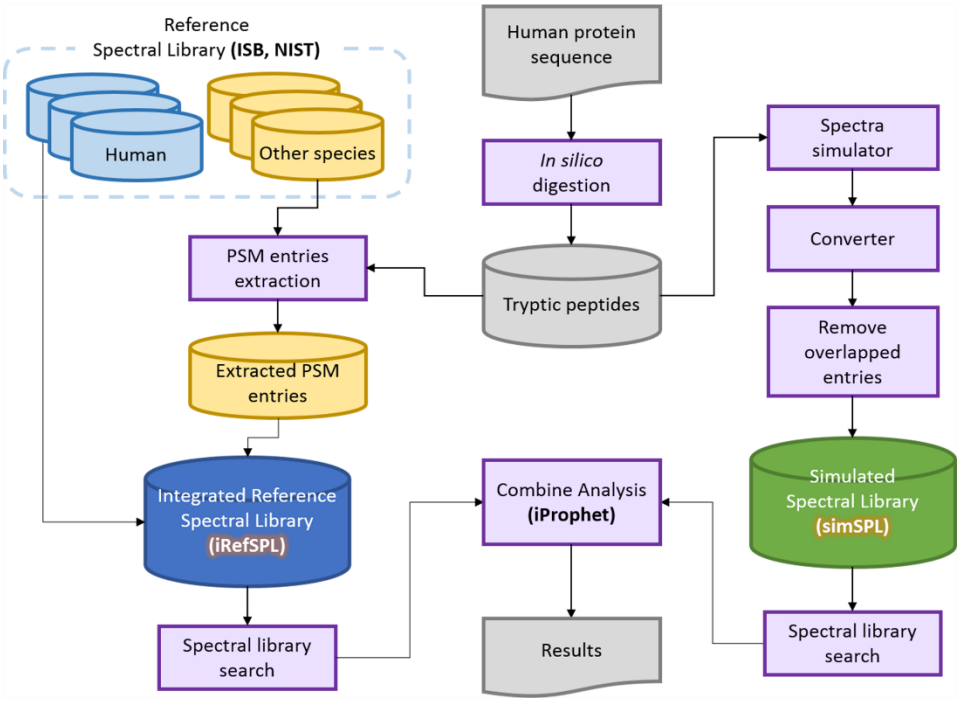



**Figure 1.** Workflow for building the integrated spectral library and multiple search results approach. Using the human tryptic peptide list, additional PSM entries were obtained from the other spectral libraries to expand the proteome coverage of the human reference spectral library called iRefSPL. We also constructed simSPL to identify novel peptides that were not covered by the iRefSPL search. In practice, the two spectral libraries were used independently in spectrum-spectrum matching and all results were combined later using iProphet.

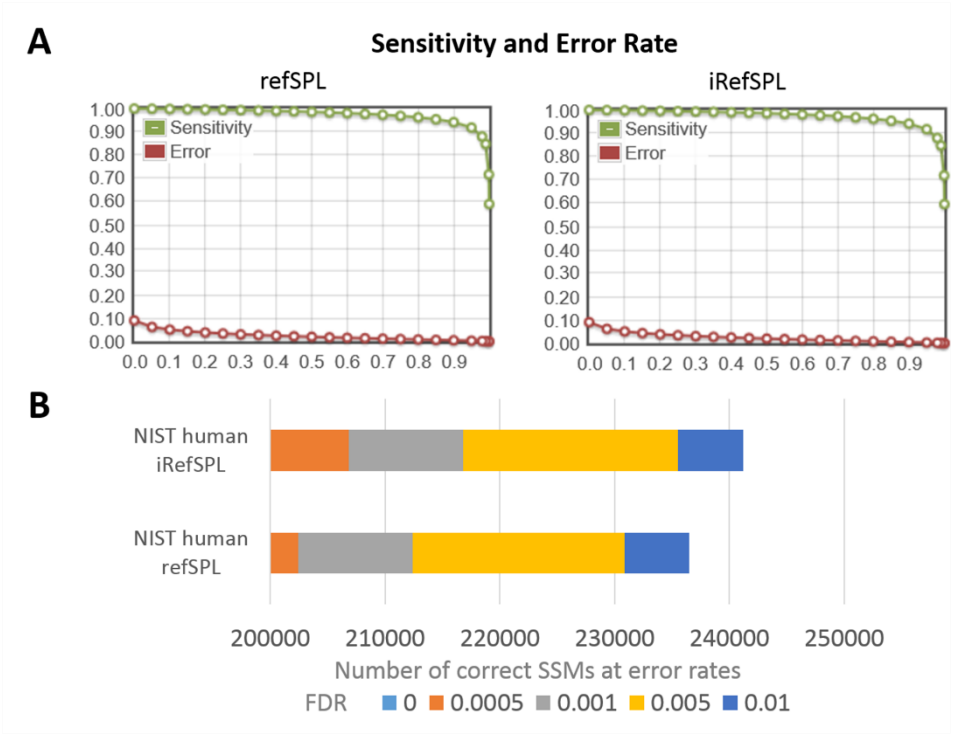



**Figure 2.** Comparison of the spectral library search results with iRefSPL and refSPL. (A) Comparison of each sensitivity and error rate model of iRefSPL and refSPL. (B) Comparison of the number of spectrum-spectrum matches through different error rates.

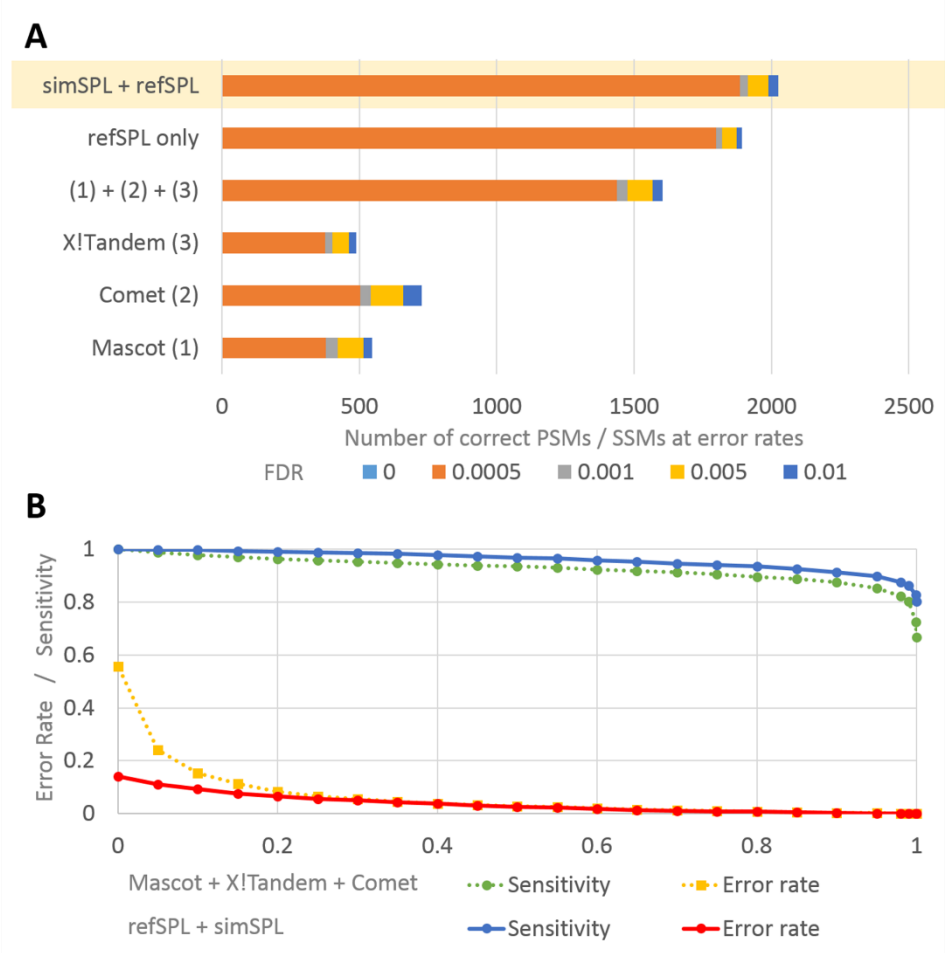



**Figure 3.** Comparison of spectral library searching with refSPL and simSPL and conventional methods for the UPS dataset analysis. (A) Comparison of matches between combination of the refSPL and simSPL, refSPL only and three sequence search engines. (B) Comparison of the sensitivity and error rates of the refSPL-simSPL combination and multiple sequence DB searching.

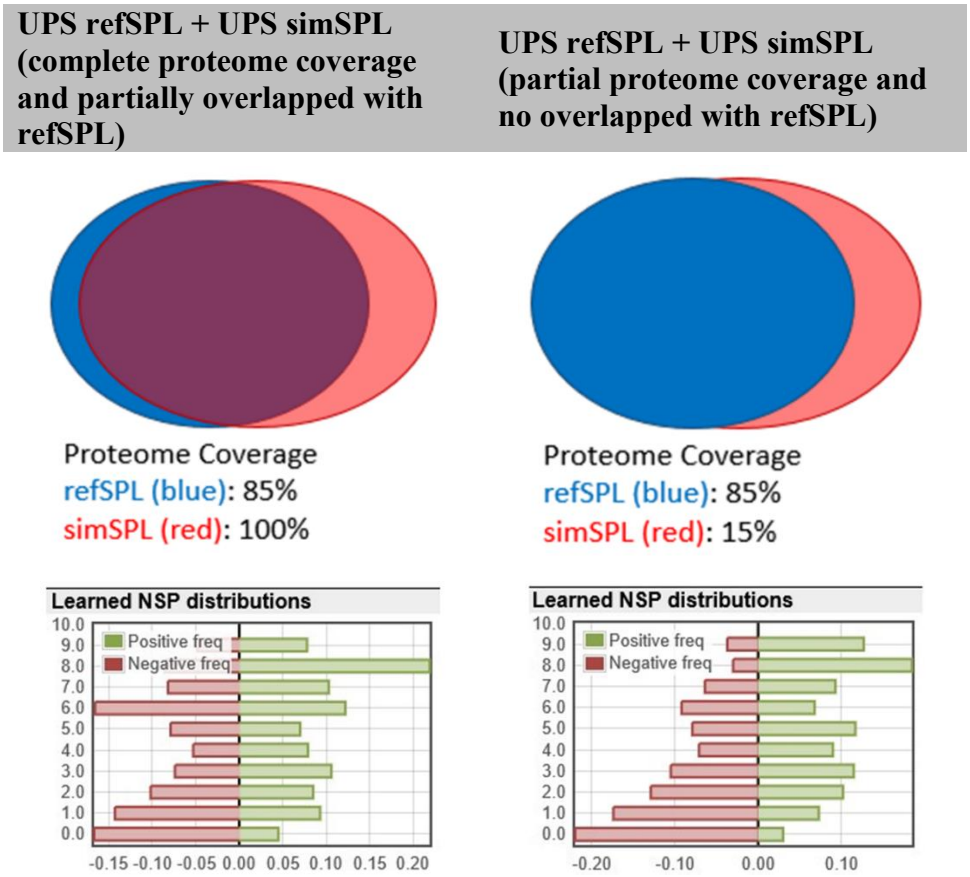



**Figure 4.** Comparison of two simSPL effect with different proteome coverages in Combo-Spec Search method.

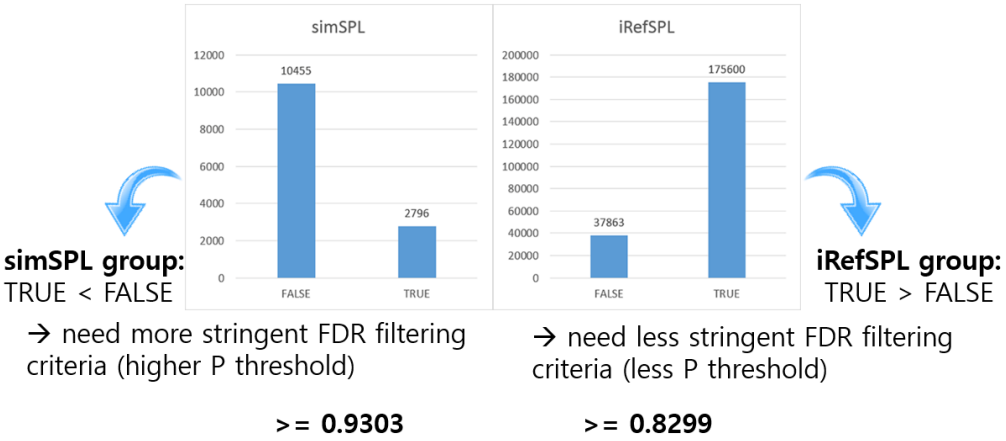



**Figure 5.** Class-Specific FDR control. The two groups of results which were processed by iRefSPL and simSPL show different true-false frequencies.

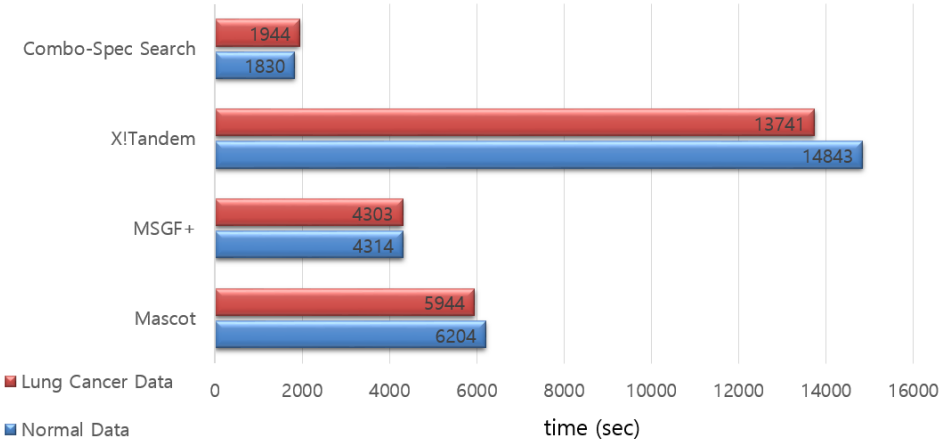



**Figure 6.** Speed of Combo-Spec Search. Combo-Spec Search shows less time to process large datasets than other sequence DB search engines.

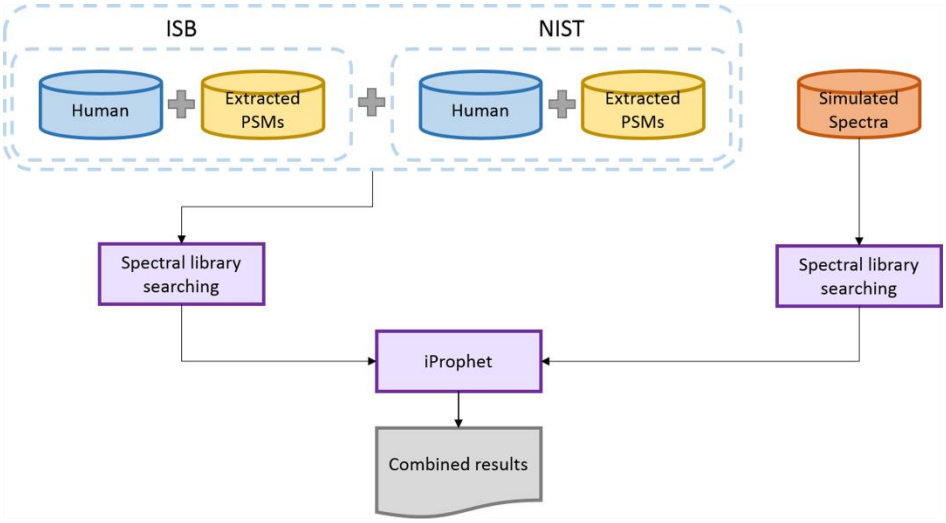



**Figure 7.** Workflow showing the human placental tissue dataset (PXD000754) analysis obtained by searching three spectral libraries and integrating the results using iProphet.

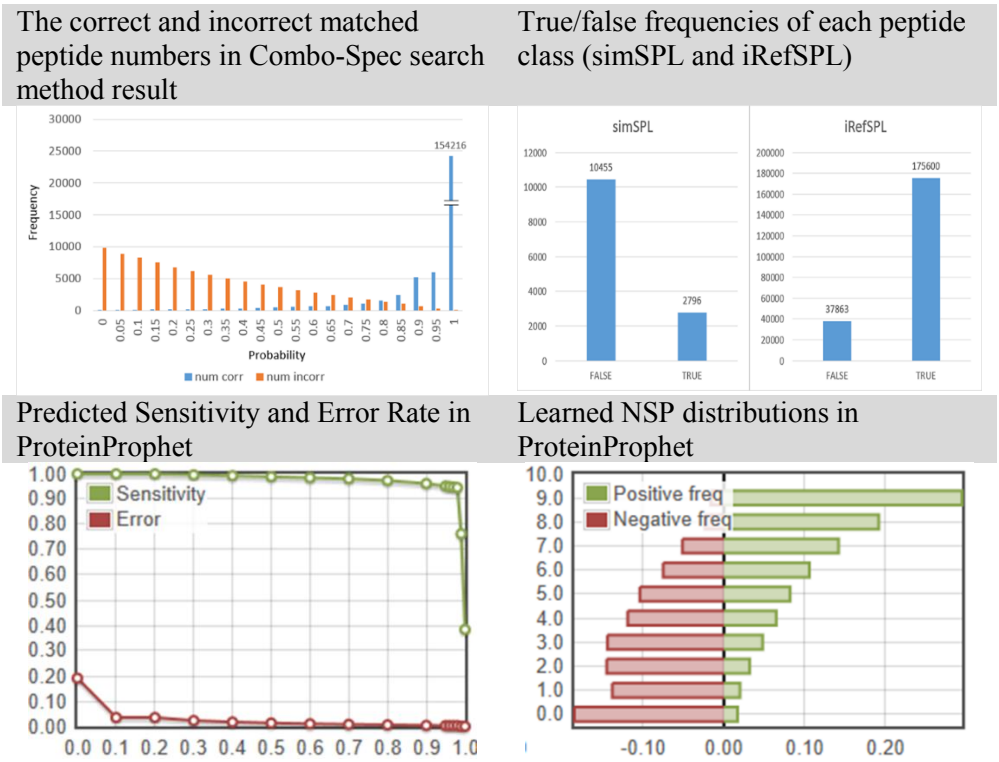
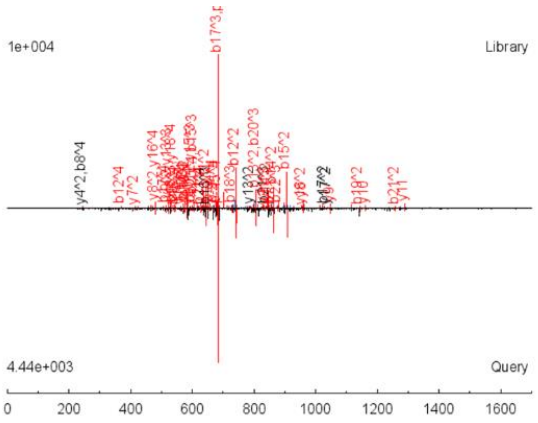
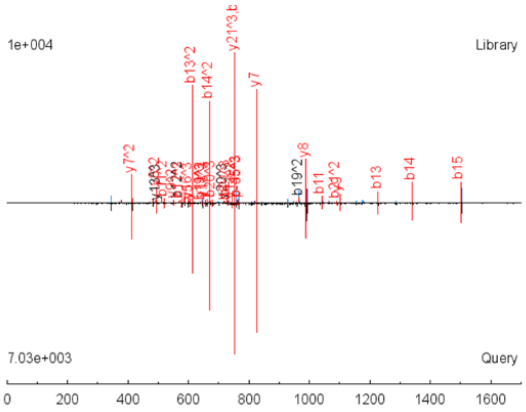



**Figure 8.** Statistics of Human placental tissue dataset.



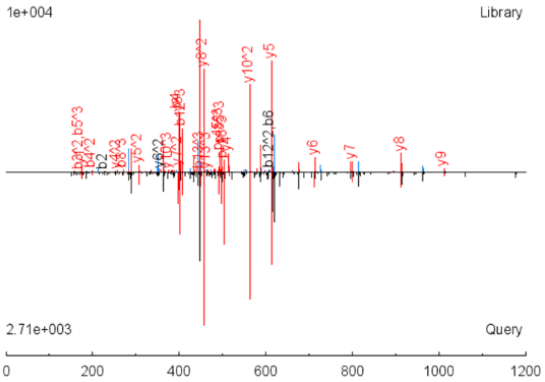

| b1+ | b2+ | b3+ | b4+ | # | AA | # | y1+ | y2+ | y3+ | y4+ |
|---|---|---|---|---|---|---|---|---|---|---|
| 116.0342 | 58.5208 | 39.3496 | 29.7640 | 1 | D | 23 | | | | |
| 253.0931 | 127.0502 | 85.0359 | 64.0287 | 2 | H | 22 | 2656.4127 | 1328.7100 | 886.1424 | 664.8586 |
| 400.1615 | 200.5844 | 134.0587 | 100.7958 | 3 | F | 21 | 2519.3537 | 1260.1805 | 840.4561 | 630.5939 |
| 513.2456 | 257.1264 | 171.7534 | 129.0669 | 4 | L | 20 | 2372.2853 | 1186.6463 | 791.4333 | 593.8538 |
| 626.3297 | 313.6685 | 209.4481 | 157.3379 | 5 | I | 19 | 2259.2013 | 1130.1043 | 753.7386 | 565.5559 |
| 697.3668 | 349.1870 | 233.1271 | 175.0972 | 6 | A | 18 | 2146.1172 | 1073.5622 | 716.0439 | 537.2948 |
| 857.3974 | 429.2024 | 286.4707 | 215.1049 | 7 | C[160] | 17 | 2075.0801 | 1038.0437 | 692.3649 | 519.5295 |
| 972.4244 | 486.7158 | 324.8130 | 243.8616 | 8 | D | 16 | 1915.0494 | 958.0284 | 639.0213 | 479.5178 |
| 1085.5084 | 543.2579 | 362.5077 | 272.1335 | 9 | L | 15 | 1800.0225 | 900.5149 | 600.6790 | 450.7611 |
| 1198.5925 | 599.7999 | 400.2023 | 300.4036 | 10 | L | 14 | 1686.9384 | 843.9729 | 562.9843 | 422.4901 |
| 1326.6511 | 663.8292 | 442.8885 | 332.4182 | 11 | Q | 13 | 1573.8544 | 787.4308 | 525.2996 | 394.2191 |
| 1482.7522 | 741.8797 | 494.9222 | 371.4435 | 12 | R | 12 | 1445.7958 | 723.4015 | 482.6035 | 362.2044 |
| 1611.7948 | 806.4010 | 537.9364 | 403.7042 | 13 | E | 11 | 1289.6947 | 645.3510 | 430.5698 | 323.1791 |
| 1725.8377 | 863.4225 | 575.9508 | 432.2149 | 14 | N | 10 | 1160.6521 | 580.8297 | 387.5556 | 290.9185 |
| 1812.8697 | 906.9385 | 604.9614 | 453.9729 | 15 | S | 9 | 1046.6092 | 523.8082 | 349.5412 | 262.4078 |
| 1941.9123 | 971.4598 | 647.9756 | 486.2335 | 16 | E | 8 | 959.5772 | 480.2922 | 320.5306 | 240.6497 |
| 2054.9964 | 1028.0018 | 685.6703 | 514.5046 | 17 | L | 7 | 830.5346 | 415.7709 | 277.5164 | 208.3891 |
| 2184.0390 | 1092.5231 | 728.6845 | 546.7652 | 18 | E | 6 | 717.4505 | 359.2289 | 239.8217 | 180.1181 |
| 2285.0866 | 1143.0470 | 762.3671 | 572.0271 | 19 | T | 5 | 588.4079 | 294.7076 | 196.8075 | 147.8574 |
| 2413.1816 | 1207.0944 | 805.0654 | 604.0509 | 20 | K | 4 | 487.3602 | 244.1838 | 163.1249 | 122.5955 |
| 2512.2500 | 1256.6286 | 838.0882 | 629.8180 | 21 | V | 3 | 359.2553 | 180.1363 | 120.4266 | 90.5718 |
| 2625.3341 | 1313.1707 | 875.7829 | 657.0890 | 22 | L | 2 | 260.1969 | 130.6021 | 87.4038 | 65.8047 |
| | | | | 23 | K | 1 | 147.1128 | 74.0600 | 49.7091 | 37.5337 |

**DHFLIAC$_{160}$DLLQRENSELETKVLK / 4** (M/Z = 693.615, P = 0.8953, iRefSPL) from sp|Q5VVM6|CCD30_HUMAN



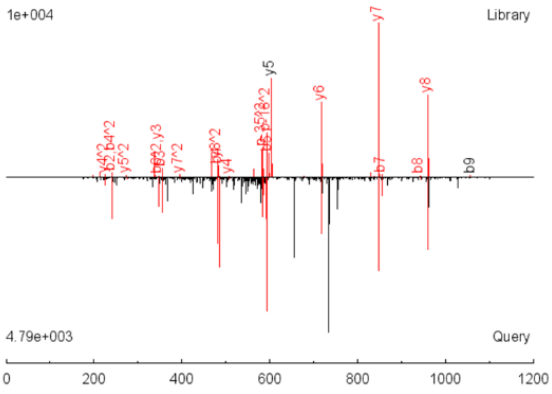

| b1+ | b2+ | b3+ | # | AA | # | y1+ | y2+ | y3+ |
|---|---|---|---|---|---|---|---|---|
| 72.0444 | 36.5258 | 24.6863 | 1 | A | 22 | | | |
| 171.1128 | 86.0600 | 57.7091 | 2 | V | 21 | 2256.1733 | 1128.5903 | 752.7293 |
| 258.1448 | 129.5761 | 86.7198 | 3 | S | 20 | 2157.1049 | 1079.0561 | 719.7065 |
| 359.1925 | 180.0999 | 120.4024 | 4 | T | 19 | 2070.0728 | 1035.5401 | 690.6958 |
| 462.2017 | 231.6045 | 154.7388 | 5 | C | 18 | 1969.0251 | 985.0162 | 657.0132 |
| 519.2232 | 260.1152 | 173.7459 | 6 | G | 17 | 1866.0160 | 933.5116 | 622.6768 |
| 590.2603 | 295.6338 | 197.4249 | 7 | A | 16 | 1808.9945 | 905.0009 | 603.6697 |
| 727.3192 | 364.1632 | 243.1112 | 8 | H | 15 | 1737.9574 | 869.4823 | 579.9906 |
| 840.4032 | 420.7053 | 280.8059 | 9 | L | 14 | 1600.8985 | 800.9529 | 534.3043 |
| 953.4873 | 477.2473 | 318.5006 | 10 | L | 13 | 1487.8144 | 744.4108 | 496.6097 |
| 1040.5193 | 520.7633 | 347.5113 | 11 | S | 12 | 1374.7304 | 687.8688 | 458.9150 |
| 1139.5877 | 570.2975 | 380.5341 | 12 | V | 11 | 1287.6983 | 644.3528 | 429.9043 |
| 1226.6198 | 613.8135 | 409.5448 | 13 | S | 10 | 1188.6299 | 594.8186 | 396.8815 |
| 1339.7038 | 670.3556 | 447.2395 | 14 | L | 9 | 1101.5979 | 551.3026 | 367.8708 |
| 1502.7672 | 751.8872 | 501.5939 | 15 | Y | 8 | 988.5138 | 494.7606 | 330.1761 |
| 1665.8305 | 833.4189 | 555.9484 | 16 | Y | 7 | 825.4505 | 413.2289 | 275.8217 |
| 1722.8520 | 861.9296 | 574.9555 | 17 | G | 6 | 662.3872 | 331.6972 | 221.4672 |
| 1819.9047 | 910.4560 | 607.3064 | 18 | P | 5 | 605.3657 | 303.1865 | 202.4601 |
| 1932.9888 | 966.9980 | 645.0011 | 19 | L | 4 | 508.3130 | 254.6601 | 170.1092 |
| 2034.0364 | 1017.5219 | 678.6837 | 20 | T | 3 | 395.2289 | 198.1181 | 132.4145 |
| 2181.1049 | 1091.0561 | 727.7065 | 21 | F | 2 | 294.1812 | 147.5942 | 98.7319 |
| | | | 22 | K | 1 | 147.1128 | 74.0600 | 49.7091 |

**AVSTCGAHLLSVSLYYGPLTFK / 3** (M/Z = 776.408, P = 0.987, iRefSPL) from sp|Q8NGV6|OR5H6_HUMAN



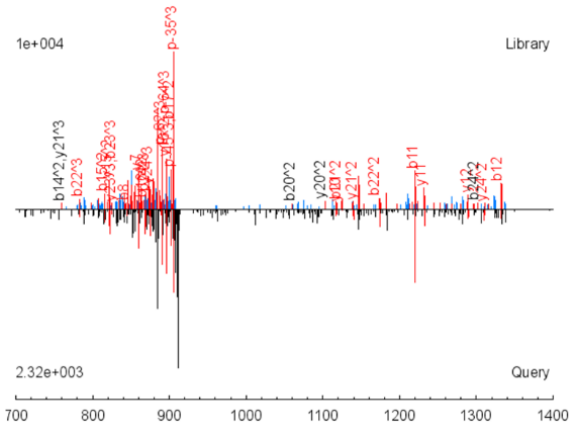

| b1+ | b2+ | b3+ | # | AA | # | y1+ | y2+ | y3+ |
|---|---|---|---|---|---|---|---|---|
| 129.0659 | 65.0366 | 43.6935 | 1 | Q | 14 | | | |
| 228.1343 | 114.5708 | 76.7163 | 2 | V | 13 | 1394.8365 | 697.9219 | 465.6170 |
| 342.1772 | 171.5922 | 114.7306 | 3 | N | 12 | 1295.7681 | 648.3877 | 432.5942 |
| 399.1987 | 200.1030 | 133.7377 | 4 | G | 11 | 1181.7252 | 591.3662 | 394.5799 |
| 512.2827 | 256.6450 | 171.4324 | 5 | L | 10 | 1124.7037 | 562.8555 | 375.5728 |
| 611.3511 | 306.1792 | 204.4552 | 6 | V | 9 | 1011.6197 | 506.3135 | 337.8781 |
| 724.4352 | 362.7212 | 242.1499 | 7 | L | 8 | 912.5513 | 456.7793 | 304.8553 |
| 811.4672 | 406.2372 | 271.1606 | 8 | S | 7 | 799.4672 | 400.2372 | 267.1606 |
| 910.5356 | 455.7715 | 304.1834 | 9 | V | 6 | 712.4352 | 356.7212 | 238.1499 |
| 1009.6040 | 505.3057 | 337.2062 | 10 | V | 5 | 613.3668 | 307.1870 | 205.1271 |
| 1122.6881 | 561.8477 | 374.9009 | 11 | L | 4 | 514.2984 | 257.6528 | 172.1043 |
| 1219.7409 | 610.3741 | 407.2518 | 12 | P | 3 | 401.2143 | 201.1108 | 134.4096 |
| 1348.7834 | 674.8954 | 450.2660 | 13 | E | 2 | 304.1615 | 152.5844 | 102.0587 |
| | | | 14 | R | 1 | 175.1190 | 88.0631 | 59.0445 |

**QVNGLVLSVVLPER/3** (M/Z = 508.303, P = 0.9919, iRefSPL) from sp|Q8IZF3|GP115_HUMAN



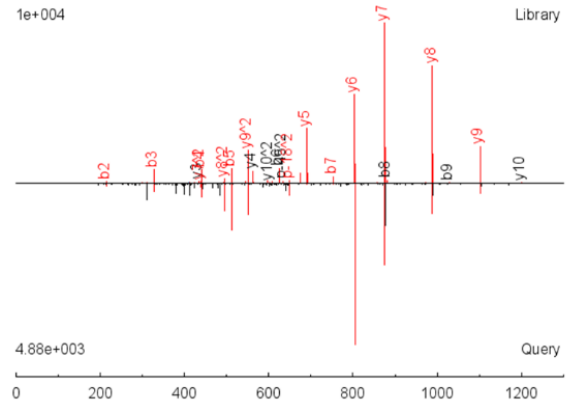

| b1+ | b2+ | # | AA | # | y1+ | y2+ |
|---|---|---|---|---|---|---|
| 129.0659 | 65.0366 | 1 | Q | 10 | | |
| 242.1499 | 121.5786 | 2 | I | 9 | 1073.5659 | 537.2866 |
| 355.2340 | 178.1206 | 3 | L | 8 | 960.4819 | 480.7446 |
| 484.2766 | 242.6419 | 4 | E | 7 | 847.3978 | 424.2026 |
| 598.3195 | 299.6634 | 5 | N | 6 | 718.3552 | 359.6813 |
| 695.3723 | 348.1898 | 6 | P | 5 | 604.3123 | 302.6598 |
| 855.4029 | 428.2051 | 7 | C[160] | 4 | 507.2595 | 254.1334 |
| 942.4349 | 471.7211 | 8 | S | 3 | 347.2289 | 174.1181 |
| 1055.5190 | 528.2631 | 9 | L | 2 | 260.1969 | 130.6021 |
| | | 10 | K | 1 | 147.1128 | 74.0600 |

**QILENPC160SLK/2** (M/Z = 601.316, P = 0.9955, iRefSPL) from sp|Q8WXK1|ASB15_HUMAN



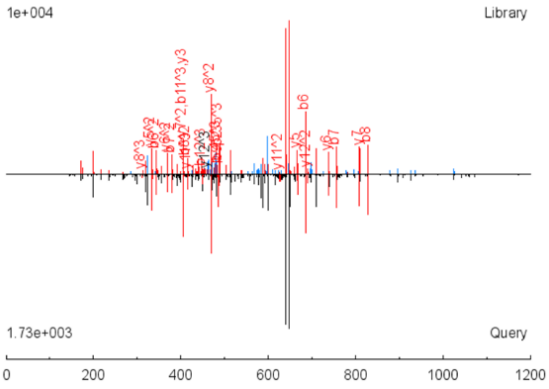

| b1+ | b2+ | b3+ | # | AA | # | y1+ | y2+ | y3+ |
|---|---|---|---|---|---|---|---|---|
| 129.0659 | 65.0366 | 43.6935 | 1 | Q | 25 | | | |
| 276.1013 | 138.5543 | 92.7053 | 2 | M[147] | 24 | 2621.3234 | 1311.1654 | 874.4460 |
| 375.1697 | 188.0885 | 125.7281 | 3 | V | 23 | 2474.2880 | 1237.6477 | 825.4342 |
| 472.2224 | 236.6149 | 158.0790 | 4 | P | 22 | 2375.2196 | 1188.1135 | 792.4114 |
| 543.2595 | 272.1334 | 181.7580 | 5 | A | 21 | 2278.1669 | 1139.5871 | 760.0605 |
| 630.2916 | 315.6494 | 210.7687 | 6 | S | 20 | 2207.1298 | 1104.0685 | 736.3814 |
| 743.3756 | 372.1915 | 248.4634 | 7 | I | 19 | 2120.0977 | 1060.5525 | 707.3708 |
| 844.4233 | 422.7153 | 282.1460 | 8 | T | 18 | 2007.0137 | 1004.0105 | 669.6761 |
| 959.4502 | 480.2288 | 320.4883 | 9 | D | 17 | 1905.9660 | 953.4866 | 635.9935 |
| 1106.4856 | 553.7465 | 369.5001 | 10 | M[147] | 16 | 1790.9390 | 895.9732 | 597.6512 |
| 1219.5697 | 610.2885 | 407.1948 | 11 | L | 15 | 1643.9036 | 822.4555 | 548.6394 |
| 1332.6538 | 666.8305 | 444.8894 | 12 | L | 14 | 1530.8196 | 765.9134 | 510.9447 |
| 1461.6964 | 731.3518 | 487.9036 | 13 | E | 13 | 1417.7355 | 709.3714 | 473.2500 |
| 1518.7178 | 759.8625 | 506.9108 | 14 | G | 12 | 1288.6929 | 644.8501 | 430.2358 |
| 1632.7607 | 816.8840 | 544.9251 | 15 | N | 11 | 1231.6715 | 616.3394 | 411.2287 |
| 1731.8232 | 866.4152 | 577.9479 | 16 | V | 10 | 1117.6286 | 559.3179 | 373.2144 |
| 1802.8663 | 901.9368 | 601.6269 | 17 | A | 9 | 1018.5601 | 509.7837 | 340.1916 |
| 1889.8983 | 945.4528 | 630.6376 | 18 | S | 8 | 947.5230 | 474.2652 | 316.5125 |
| 2002.9824 | 1001.9948 | 668.3323 | 19 | I | 7 | 860.4910 | 430.7491 | 287.5018 |
| 2116.0664 | 1058.5369 | 706.0270 | 20 | L | 6 | 747.4069 | 374.2071 | 249.8072 |
| 2245.1090 | 1123.0581 | 749.0412 | 21 | E | 5 | 634.3229 | 317.6651 | 212.1125 |
| 2344.1774 | 1172.5923 | 782.0640 | 22 | V | 4 | 505.2803 | 253.1438 | 169.0983 |
| 2475.2179 | 1238.1126 | 825.7442 | 23 | M | 3 | 406.2119 | 203.6096 | 136.0755 |
| 2603.2765 | 1302.1419 | 868.4303 | 24 | Q | 2 | 275.1714 | 138.0893 | 92.3953 |
| | | | 25 | K | 1 | 147.1128 | 74.0600 | 49.7091 |

**QM$_{147}$VPASITDM$_{147}$LLEGNVASILEVMQK/3** (M/Z = 917.132, P = 0.9545, iRefSPL) from sp|Q8NE28|SGK71_HUMAN



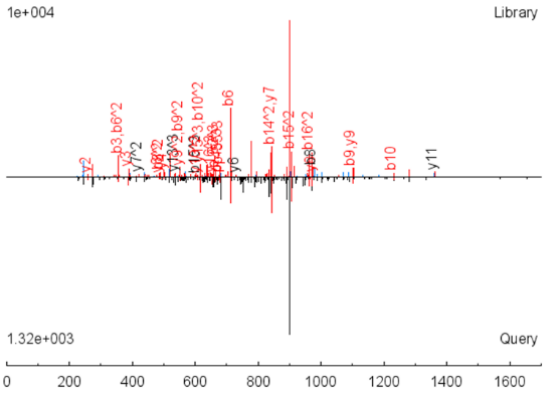

| b1+ | b2+ | # | AA | # | y1+ | y2+ |
|---|---|---|---|---|---|---|
| 116.0342 | 58.5208 | 1 | D | 11 | | |
| 215.1026 | 108.0550 | 2 | V | 10 | 1200.7133 | 600.8603 |
| 328.1867 | 164.5970 | 3 | I | 9 | 1101.6449 | 551.3261 |
| 441.2707 | 221.1390 | 4 | L | 8 | 988.5608 | 494.7840 |
| 512.3079 | 256.6576 | 5 | A | 7 | 875.4767 | 438.2420 |
| 625.3919 | 313.1996 | 6 | I | 6 | 804.4396 | 402.7235 |
| 753.4505 | 377.2289 | 7 | Q | 5 | 691.3556 | 346.1814 |
| 881.5091 | 441.2582 | 8 | Q | 4 | 563.2970 | 282.1521 |
| 1028.5445 | 514.7759 | 9 | M[147] | 3 | 435.2384 | 218.1228 |
| 1141.6285 | 571.3179 | 10 | I | 2 | 288.2030 | 144.6051 |
| | | 11 | R | 1 | 175.1190 | 88.0631 |

**DVILAIQQM$_{147}$IR/2** (M/Z = 658.374, P = 0.9987, simSPL) from sp|Q6IEU7|OR5MA_HUMAN



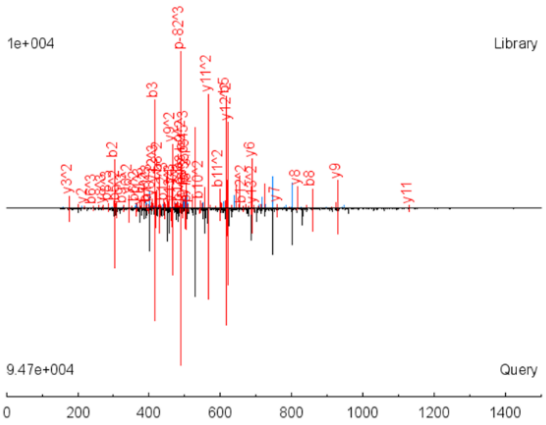

| b1+ | b2+ | b3+ | # | AA | # | y1+ | y2+ | y3+ |
|---|---|---|---|---|---|---|---|---|
| 116.0342 | 58.5208 | 39.3496 | 1 | D | 13 | | | |
| 244.0928 | 122.5500 | 82.0358 | 2 | Q | 12 | 1377.7559 | 689.3816 | 459.9235 |
| 373.1354 | 187.0713 | 125.0500 | 3 | E | 11 | 1249.6973 | 625.3523 | 417.2373 |
| 444.1725 | 222.5899 | 148.7290 | 4 | A | 10 | 1120.6547 | 560.8310 | 374.2231 |
| 557.2566 | 279.1319 | 186.4237 | 5 | L | 9 | 1049.6176 | 525.3124 | 350.5440 |
| 685.3151 | 343.1612 | 229.1099 | 6 | Q | 8 | 936.5335 | 468.7704 | 312.8494 |
| 756.3523 | 378.6798 | 252.7889 | 7 | A | 7 | 808.4749 | 404.7411 | 270.1632 |
| 827.3894 | 414.1983 | 276.4680 | 8 | A | 6 | 737.4378 | 369.2226 | 246.4841 |
| 974.4578 | 487.7325 | 325.4908 | 9 | F | 5 | 666.4007 | 333.7040 | 222.8051 |
| 1087.5418 | 544.2746 | 363.1855 | 10 | L | 4 | 519.3323 | 260.1698 | 173.7823 |
| 1215.6368 | 608.3220 | 405.8838 | 11 | K | 3 | 406.2482 | 203.6278 | 136.0876 |
| 1346.6773 | 673.8423 | 449.5639 | 12 | M | 2 | 278.1533 | 139.5803 | 93.3893 |
| | | | 13 | K | 1 | 147.1128 | 74.0600 | 49.7091 |

**DQEALQAAFLKMK/3** (M/Z = 498.266, P = 0.9949, iRefSPL) from sp|O75343|GCYB2_HUMAN



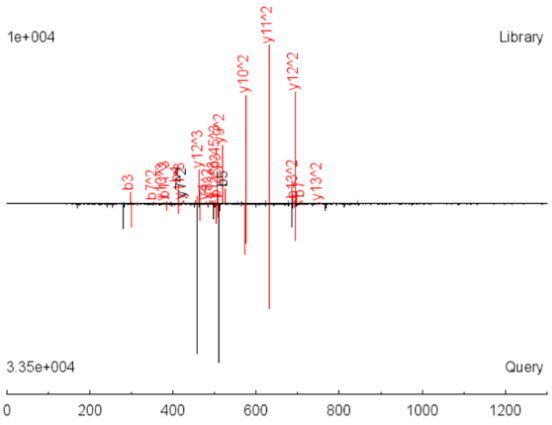

| b1+ | b2+ | b3+ | # | AA | # | y1+ | y2+ | y3+ |
|---|---|---|---|---|---|---|---|---|
| 116.0342 | 58.5208 | 39.3496 | 1 | D | 17 | | | |
| 229.1183 | 115.0628 | 77.0443 | 2 | L | 16 | 1958.0990 | 979.5532 | 653.3712 |
| 357.1769 | 179.0921 | 119.7305 | 3 | Q | 15 | 1845.0150 | 923.0111 | 615.6765 |
| 470.2609 | 235.6341 | 157.4252 | 4 | L | 14 | 1716.9564 | 858.9818 | 572.9903 |
| 598.3450 | 292.1761 | 195.1198 | 5 | L | 13 | 1603.8723 | 802.4398 | 535.2956 |
| 712.3876 | 356.6974 | 238.1340 | 6 | E | 12 | 1490.7883 | 745.8978 | 497.6009 |
| 859.4230 | 430.2151 | 287.1458 | 7 | M[147] | 11 | 1361.7457 | 681.3765 | 454.5867 |
| 973.4659 | 487.2366 | 325.1602 | 8 | N | 10 | 1214.7103 | 607.8588 | 405.5749 |
| 1101.5609 | 551.2841 | 367.8585 | 9 | K | 9 | 1100.6674 | 550.8373 | 367.5606 |
| 1230.6035 | 615.8054 | 410.8727 | 10 | E | 8 | 972.5724 | 486.7898 | 324.8623 |
| 1344.6464 | 672.8268 | 448.8870 | 11 | N | 7 | 843.5298 | 422.2685 | 281.8481 |
| 1473.6890 | 737.3481 | 491.9012 | 12 | E | 6 | 729.4869 | 365.2471 | 243.8338 |
| 1572.7574 | 786.8823 | 524.9240 | 13 | V | 5 | 600.4443 | 300.7258 | 200.8196 |
| 1685.8414 | 843.4244 | 562.6187 | 14 | L | 4 | 501.3759 | 251.1916 | 167.7968 |
| 1813.9364 | 907.4718 | 605.3170 | 15 | K | 3 | 388.2918 | 194.6495 | 130.1021 |
| 1927.0205 | 964.0139 | 643.0117 | 16 | I | 2 | 260.1969 | 130.6021 | 87.4038 |
| | | | 17 | K | 1 | 147.1128 | 74.0600 | 49.7091 |

**DLQLLEM$_{147}$NKENEVLKIK/3** (M/Z = 691.714, P = 0.9721, iRefSPL) from sp|Q9H2F9|CCD68_HUMAN



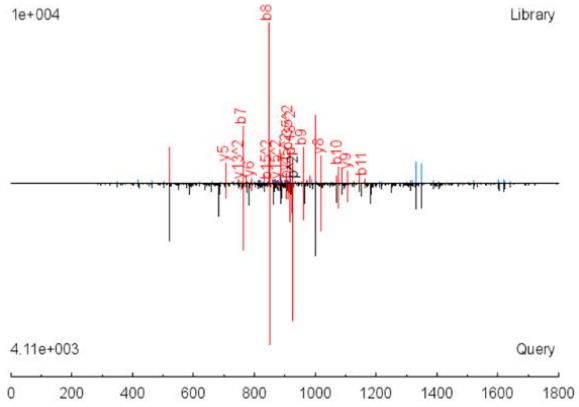

| $b^{1+}$ | $b^{2+}$ | $b^{3+}$ | # | AA | # | $y^{1+}$ | $y^{2+}$ | $y^{3+}$ |
|---|---|---|---|---|---|---|---|---|
| 157.1084 | 79.0578 | 53.0410 | 1 | R | 14 | | | |
| 304.1438 | 152.5755 | 102.0528 | 2 | M[147] | 13 | 1389.6202 | 695.3137 | 463.8783 |
| 417.2279 | 209.1176 | 139.7475 | 3 | L | 12 | 1242.5848 | 621.7960 | 414.8665 |
| 546.2704 | 273.6389 | 182.7617 | 4 | E | 11 | 1129.5007 | 565.2540 | 377.1718 |
| 617.3076 | 309.1574 | 206.4407 | 5 | A | 10 | 1000.4582 | 500.7327 | 334.1576 |
| 730.3916 | 365.6994 | 244.1354 | 6 | L | 9 | 929.4210 | 465.2142 | 310.4785 |
| 787.4131 | 394.2102 | 263.1425 | 7 | G | 8 | 816.3370 | 408.6721 | 272.7838 |
| 858.4502 | 429.7287 | 286.8216 | 8 | A | 7 | 759.3155 | 380.1614 | 253.7767 |
| 987.4928 | 494.2500 | 329.8358 | 9 | E | 6 | 688.2784 | 344.6428 | 230.0977 |
| 1084.5455 | 542.7764 | 362.1867 | 10 | P | 5 | 559.2358 | 280.1216 | 187.0835 |
| 1198.5885 | 599.7979 | 400.2010 | 11 | N | 4 | 462.1831 | 231.5952 | 154.7325 |
| 1327.6311 | 664.3192 | 443.2152 | 12 | E | 3 | 348.1401 | 174.5737 | 116.7182 |
| 1456.6736 | 728.8405 | 486.2294 | 13 | E | 2 | 219.0975 | 110.0524 | 73.7040 |
| | | | 14 | A | 1 | 90.0549 | 45.5311 | 30.6898 |

**RM$_{147}$LEALGAEPNEEA/3** (M/Z = 515.912, P = 0.9683, iRefSPL) from tr|C9J6K1|C9J6K1_HUMAN



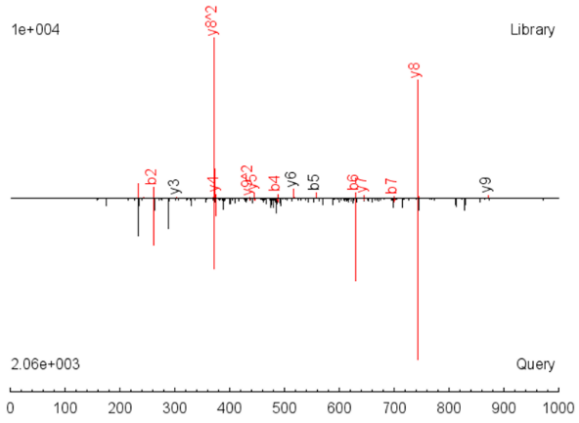

| $b^{1+}$ | $b^{2+}$ | $b^{3+}$ | # | AA | # | $y^{1+}$ | $y^{2+}$ | $y^{3+}$ |
|---|---|---|---|---|---|---|---|---|
| 58.0287 | 29.5180 | 20.0144 | 1 | G | 14 | | | |
| 171.1128 | 86.0600 | 57.7091 | 2 | L | 13 | 1502.8471 | 751.9272 | 501.6206 |
| 299.1714 | 150.0893 | 100.3953 | 3 | Q | 12 | 1389.7631 | 695.3852 | 463.9259 |
| 412.2554 | 206.6314 | 138.0900 | 4 | I | 11 | 1261.7045 | 631.3559 | 421.2397 |
| 525.3395 | 263.1734 | 175.7847 | 5 | L | 10 | 1148.6204 | 574.8139 | 383.5450 |
| 582.3610 | 291.6841 | 194.7918 | 6 | G | 9 | 1035.5364 | 518.2718 | 345.8503 |
| 710.4195 | 355.7134 | 237.4780 | 7 | Q | 8 | 978.5149 | 489.7611 | 326.8432 |
| 811.4672 | 406.2373 | 271.1606 | 8 | T | 7 | 850.4563 | 425.7318 | 284.1570 |
| 924.5513 | 462.7793 | 308.8553 | 9 | L | 6 | 749.4087 | 375.2080 | 250.4744 |
| 1080.6524 | 540.8298 | 360.8890 | 10 | R | 5 | 636.3246 | 318.6659 | 212.7797 |
| 1151.6895 | 576.3484 | 384.5680 | 11 | A | 4 | 480.2235 | 240.6154 | 160.7460 |
| 1238.7215 | 619.8644 | 413.5787 | 12 | S | 3 | 409.1864 | 205.0968 | 137.0670 |
| 1385.7569 | 693.3821 | 462.5905 | 13 | M[147] | 2 | 322.1544 | 161.5808 | 108.0563 |
| | | | 14 | R | 1 | 175.1190 | 88.0631 | 59.0445 |

**GLQILGQTLRASM147R/3** (M/Z = 520.628, P = 0.9957 , simSPL) from sp|Q96RP8|KCNA7_HUMAN




| $b^{1+}$ | $b^{2+}$ | # | AA | # | $y^{1+}$ | $y^{2+}$ |
|---|---|---|---|---|---|---|
| 115.0502 | 58.0287 | 1 | N | 16 | | |
| 243.1452 | 122.0762 | 2 | K | 15 | 1752.8771 | 876.9422 |
| 356.2292 | 178.6183 | 3 | L | 14 | 1624.7822 | 812.8947 |
| 443.2613 | 222.1343 | 4 | S | 13 | 1511.6981 | 756.3527 |
| 546.2705 | 273.6389 | 5 | C | 12 | 1424.6661 | 712.8367 |
| 649.2796 | 325.1435 | 6 | C | 11 | 1321.6569 | 661.3321 |
| 762.3637 | 381.6855 | 7 | I | 10 | 1218.6477 | 609.8275 |
| 849.3957 | 425.2015 | 8 | S | 9 | 1105.5636 | 553.2855 |
| 962.4798 | 481.7435 | 9 | I | 8 | 1018.5316 | 509.7694 |
| 1075.5639 | 538.2856 | 10 | I | 7 | 905.4475 | 453.2274 |
| 1162.5959 | 581.8016 | 11 | S | 6 | 792.3635 | 396.6854 |
| 1299.6548 | 650.3310 | 12 | H | 5 | 705.3315 | 353.1694 |
| 1428.6974 | 714.8523 | 13 | E | 4 | 568.2726 | 284.6399 |
| 1591.7607 | 796.3840 | 14 | Y | 3 | 439.2300 | 220.1186 |
| 1692.8084 | 846.9078 | 15 | T | 2 | 276.1666 | 138.5870 |
| | | 16 | R | 1 | 175.1190 | 88.0631 |

**NKLSCCISIISHEYTR/2** (M/Z = 933.964, P = 0.9243, iRefSPL) from sp|Q8N687|DB125_HUMAN




| $b^{1+}$ | $b^{2+}$ | # | AA | # | $y^{1+}$ | $y^{2+}$ |
|---|---|---|---|---|---|---|
| 132.0478 | 66.5275 | 1 | M | 10 | | |
| 261.0904 | 131.0488 | 2 | E | 9 | 871.4268 | 436.2170 |
| 358.1431 | 179.5752 | 3 | P | 8 | 742.3842 | 371.6957 |
| 487.1857 | 244.0965 | 4 | E | 7 | 645.3314 | 323.1694 |
| 558.2228 | 279.6151 | 5 | A | 6 | 516.2889 | 258.6481 |
| 629.2599 | 315.1336 | 6 | A | 5 | 445.2517 | 223.1295 |
| 700.2970 | 350.6522 | 7 | A | 4 | 374.2146 | 187.6110 |
| 757.3185 | 379.1629 | 8 | G | 3 | 303.1775 | 152.0924 |
| 828.3556 | 414.6814 | 9 | A | 2 | 246.1561 | 123.5817 |
| | | 10 | R | 1 | 175.1190 | 88.0631 |

**MEPEAAAGAR/2** (M/Z = 501.737, P = 0.979, simSPL) from sp|P57055|DSCR6_HUMAN

**Figure 9.** The spectrum view and matched peaks of the newly identified missing proteins

# SUBJECT II

## Epsilon-Q: An Automated Analyzer Interface for Mass Spectral Library Search and Label-free Quantification

# 1. Introduction

Mass spectrometry (MS) is a widely used proteome analytical tool for biomedical science. Proteins in sample mixtures can be detected and quantified by MS-coupled high performance liquid chromatography (HPLC) in a high-throughput approach. Because of rapid advances in MS instruments, experimental methods, and computing power, low-abundance proteins present in biological and clinical samples can now be detected and quantified with high levels of accuracy in a short time (Craig, et al., 2004; Eng, et al., 1994; Liu, et al., 2004; Perkins, et al., 1999). Because proteins are macromolecule, they are fragmented into peptides by enzyme digestion (Aebersold, et al., 2003; Chait, 2006) before analysis by mass spectrometry. A bottom-up proteomic approach makes it possible to analyze these peptides by MS. Many peptides derive from a single protein and can be separated by HPLC coupled with MS. This technique is called “shotgun proteomics” (Washburn, *et al.*, 2001; Wolters, *et al.*, 2001).

Protein sequence database (DB) searching is a widely-used method for matching and assigning peptide sequences to mass spectra. SEQUEST (Eng, et al., 1994), X!Tandem (Fenyo, et al., 2003), Comet (Eng, *et al.*, 2013), Mascot (Perkins, et al., 1999) and MS-GF+ (Kim, et al., 2014) are widely used sequence DB search tools. These tools produce proteolytic peptide lists and calculate fragment ion m/z values according to specific charge states and modifications using a reference protein sequence DB. Mass spectra can usually be analyzed by a sequence DB search method by preparing an appropriate sequence DB and inputting suitable parameters (Steen, et al., 2004; Zhang, et al., 2014). However, a large and complicated input dataset, sequence DB and

modification of multiple variable options can cause extended processing times. In particular, false-positive errors may arise because these methods only use m/z values in sequence-to-spectrum matching (Yen, et al., 2011).

An alternative sequencing method is spectral library searching (Yates, et al., 1998). Peptide spectral libraries contain curated, annotated, and unique peptide sequences for tandem mass spectrum pairing. The peptide-to-spectrum matches (PSMs) are used as a template to identify peptide sequences in experimental spectra. Because this method uses curated PSMs for spectrum-to-spectrum matching, it provides more sensitive and accurate results in a given time than sequence DB searching method (Craig, et al., 2006; Lam, et al., 2007; Yen, et al., 2011). Some peptide spectral libraries are publicly available. The National Institute of Standards and Technology (NIST) and PeptideAtlas, operated by the Institute for Systems Biology (ISB), are representative public peptide spectral library providers. Some researchers and laboratories build customized peptide libraries for their studies (Lam, 2011; Lam, et al., 2008). These peptide libraries, built by different institutes and researchers, are generated with custom criteria and false-positive entries. Because of this, merging spectral libraries from different sources can cause an increase in the FDR (Deutsch, *et al.*, 2015b). For this reason, it is recommended that spectral library searching be independently performed, and thus FDR can be estimated for that peptide library. In addition, peptide spectral libraries only contain known peptides, so this method has limitations if used to find novel and previously unobserved peptide sequences (Yen, et al., 2011; Yen, et al., 2009). To overcome such limitations, we previously designed the Combo-Spec Search method, using public or lab-based curated spectral libraries and simulated spectral libraries (simSPLs) to fill gaps in proteome coverage (Cho, *et al.*, 2015). This method provides improved sensitivity and expended proteome coverage, however,

two or more spectral libraries are needed to conduct a search for a single MS dataset. It is, therefore, a cumbersome and tedious task compared to the sequence DB search method.

Quantitative comparison of proteome expression is a challenging issue in disease-related proteome research and biomarker discovery. The label-free quantitation method is particularly suitable for quantitative analyses by MS (James, 1997). This method directly uses the peak signal intensity of the extracted ion chromatogram (XIC) or the spectral count to estimate peptide or protein abundance. In general, the peak intensity is influenced by peptide ionization efficiencies and chemical environments, indicating that the sensitivity of mass spectrometry varies between peptides. Hence, we can overcome the limitations of the labeling method, which requires a complex sample pre-processing step and limits the number of samples in each experiment (Bantscheff, *et al.*, 2007; Chelius, *et al.*, 2002; Lill, 2003; Wang, *et al.*, 2003).

Spectral counting is used to determine protein abundance, based on the number of spectra matched to its peptides. Even though it is conceptually simple, spectral counting must be sensitive enough to estimate the relative protein abundance ratio over a large dynamic range. However, it sometimes generates false estimates for low-abundance proteins because spectral counting assigns an equivalent value of 1 for each spectrum of a peptide (Fu, *et al.*, 2008; Ishihama, *et al.*, 2005; Liu, et al., 2004; Old, *et al.*, 2005). The XIC allows comparison of the peak areas between peptides. This process is simple, and shows linearity in comparing peptide or protein abundance. Recently, several studies have demonstrated that the XICs of selected peptide ions correlate well with protein abundance in large or complex biological samples. However, the selection and differentiation of peptide peak areas from neighboring peaks are often difficult. Therefore, this problem must be resolved for successful quantitative analyses using XIC (Chelius, et

al., 2002).

In this study, we have developed a new software package called Epsilon-Q, which supports Combo-Spec Search and label-free quantification methods. This software supports standard MS data formats. Epsilon-Q allows automatic indexing, multiple spectral library searching and calculation of the sum of precursor ion peak intensities for user-generated datasets. Epsilon-Q also supports multi-thread processing, which enables to multiple data files to be processed concurrently. We set Epsilon-Q system by combining this automatic interface software with Combo-Spec Search method and analyzed controlled datasets with various degrees of biological complexity. With a user-friendly graphical interface, Epsilon-Q system demonstrates good performance in the identification and quantitative analysis of proteins. We anticipate that Epsilon-Q system will help users to achieve improved detectability when identifying proteins, and perform comparative analyses of biological samples.

# 2. Materials and Methods

## 2.1. Benchmark Datasets

To evaluate the performance of Epsilon-Q system, datasets containing known ratios and dynamic quantitative ranges are required. The universal proteomics standard (UPS) (Sigma-Aldrich, St. Louse, MO. USA) is a standard protein mixture containing 48 purified human recombinant proteins. UPS1 is composed of 5 pmol of each of the 48 proteins. UPS2 contains the same proteins as UPS1, however, the amount of each protein ranges from 0.5 fmol to 50 pmol. We obtained three UPS1 and UPS2 datasets from the ProteomeXchange repository (Table 1).

## 2.2. Peptide and Protein Identification

Raw MS data files were converted to .mgf and .mzML formats for each search engine using MSConvertGUI (ProteoWizard)(Chambers, *et al.*, 2012). For Epsilon-Q, the conversion parameters were as follows: 32-bit binary encoding precision, and "peak picking" filter. We prepared protein sequence DBs which included the 48 UPS proteins and *E.coli* proteins. All sequence DBs were obtained from UniProt DB. To perform the sequence DB searches, we used Mascot server v2.5, X!Tandem (2013.06.15.1 – LabKey, Insilicos, ISB) and Andromeda, built-in to MaxQuant (Cox, *et al.*, 2008) v1.5.2.8. The sequence spectral library search parameters used were: trypsin for protein digestion, carbamidomethylation at cysteine residues for fixed modifications, oxidation at methionine

and acetylation at protein N-terminal residues for variable modifications, a maximum of two missed cleavages, 5 ppm MS tolerance, and 0.6 Da MS/MS tolerance. Two charge states, 2+ and 3+, were considered. Peptide spectral libraries for UPS, *E.coli* and yeast cell lysate were obtained from the NIST peptide library repository. We built simSPLs for UPS, *E.coli* and yeast proteins using protein sequence DBs, as described previously. Tryptic peptides 7 to 45 amino acids in length, and with a maximum of 2 missed cleavage sites, were prepared. MassAnalyzer (Zhang, 2004; 2005) (version 2.3.1) was used to simulate the MS/MS spectra of the selected peptides using the following simulation parameters: Orbitrap Velos instrument profile with CID fragmentation mode, isolation width of 2.5, resolution of 800, collision energy (V) of 35, and activation time of 30 ms. We considered three charge states, 2+ to 4+, precursors, and added additional spectra which had two types of modification by the semi-empirical modification method (Hu, et al., 2011; Suni, *et al.*, 2015b): carbamidomethylation at cysteine residues for fixed modifications, and oxidation at methionine residues for variable modifications. The simulated spectra were converted to the *.splib format using SpectraST (Lam, et al., 2007), and all peptide-to-spectrum matches already included in the reference spectral library (refSPL; composed of annotated experimental spectra or publicly available spectral libraries) were removed. Spectral library searches were performed using SpectraST v5.0. All results were filtered to achieve a false-positive error rate of less than 1% for each peptide and protein.

## 2.3. Statistical Estimation and Result Integration

The results of each search were analyzed using PeptideProphet and ProteinProphet (built in Trans-Proteome Pipeline version 4.8.0)(Deutsch, *et al.*, 2015a)

with default parameters. We used decoy hits and a non-parametric model to determine the negative frequency, and determined two-peptide probability thresholds by class-specific FDR filtering (Nesvizhskii, 2014). Each threshold was established by separate FDR estimations in two classes. Peptide and protein hit probability score thresholds were determined by FDR estimation. All protein and peptide hits were filtered and parsed by the predetermined thresholds.

## 2.4. Quantification and Removal of Outliers

Peptide hits having a higher probabilistic score than the threshold were selected to calculate the sum of the precursor peak intensities. Using the precursor peak information, nearby peaks were scanned to find the maximum peak. Based on the maximum peak information, Epsilon-Q sequentially scans precursor peak groups and isotopic peaks. All candidate peaks detected were grouped into a feature. These precursor peaks groups were then used to calculate the peptide abundance indexes, as the sum of the peak intensities. For each protein, peptide abundance ratios were calculated using sample pairs and estimated outliers. The outlier detection was performed by median absolute deviation (MAD)(Rousseeuw, *et al.*, 1993).

# 3. Results

## 3.1. Epsilon-Q Workflow

Epsilon-Q supports searches of multiple peptide spectral libraries for user-generated mass spectrum datasets (Figure 1) that have been used in the "Combo-Spec Search" method. This is designed to overcome a lack of proteome coverage in a peptide spectral library. If a user wishes to find specific sequences which are not included in a public spectral library, simSPL searching may provide a way to detect those sequences. Because the use of multiple spectral libraries generates duplicate results in each mass spectral dataset, refining these searches can be a time-consuming and cumbersome task. To improve the efficiency of this process, Epsilon-Q is designed to support multi-thread functional to process results in parallel. Each result is statistically evaluated by PeptideProphet and ProteinProphet, and filtered by its FDR. After the results are combined, the sum of the intensities of the precursor ion group is calculated for each peptide. Epsilon-Q calculates peptide abundance ratios and detects outliers for each protein. The sums of peptide abundances are calculated without the outliers, as protein abundance indexes. The results are saved as a text-based file (csv format).

## 3.2. SimSPL Builder Features and Workflow

SimSPL Builder supports features for building simSPLs to overcome the lack of proteome coverage in peptide spectral libraries. First, using a protein sequence DB

(FASTA format), SimSPL Builder creates a tryptic peptide list. In the next step, MassAnalyzer is used to simulate a tandem mass spectrum for a given peptide set. MassAnalyzer provides various MS instrument profiles, such as LTQ, Q-TOF, Orbitrap, and Q Exactive. SimSPL Builder converts the simulated tandem mass spectrum to splib format, so it can be used SpectraST. SimSPL Builder also provides as interface to add semi-empirical modifications and decoy generation to false-positive estimates (Figure 2).

## 3.3. Precursor Ion Peak Detection

Epsilon-Q sequentially processes all the candidate peaks around the peak of maximum intensity. Figure 3 shows the precursor peak detection workflow in Epsilon-Q. First, Epsilon-Q roughly scans the local peaks to find the peak with the maximum intensity, higher than a given threshold (m/z and time window). After the selection of the maximum peak, two-way candidate peak detection is performed. Precursor peaks are detected along the retention time axis, and isotope peaks are detected along the m/z axis (Figure 3B). Isotope peak detection is based on the m/z threshold and the lower and upper isolation window offset. This offset is automatically determined by Epsilon-Q based on the raw MS data files (Figure 3C). The scan time ranges used to find precursor peaks are determined based on the extracted ion peak width-at-half-height. This process is performed for each assigned precursor peak and all detected isotopic peaks (Figure 3D).

## 3.4. Estimation of Peptide and Protein Detection Performance by Epsilon-Q

The sensitivity of Epsilon-Q for protein and peptide detection was tested using the UPS1 in yeast datasets (Lamus, C *et al*., PXD001819(Ramus, *et al.*, 2016)). The datasets were composed of UPS1 sets that were spiked at different concentrations into yeast cell lysate. Figure 4 shows the Epsilon-Q detection performance, as compared with three popular sequence DB search engines. At the protein level, Epsilon-Q yielded more identified proteins than other methods, particularly those present at low concentrations (Figure 4A). For samples containing 5 fmol of each protein, Epsilon-Q shows more distinct peptide sequences compared with other tools (Figure 4B). Figure 5 shows sequence-to-spectrum matches by Epsilon-Q at a concentration of 5 fmol. Peptides having novel sequences for each protein were not detected by any of the other search engines (X!Tandem, Mascot, or MaxQuant). These results show that Epsilon-Q has advantages over other search engines in the detection of peptides and proteins, especially those present at low concentrations.

## 3.5. Estimation of Quantitative Performance by Epsilon-Q

To estimate the quantitative performance of Epsilon-Q, we prepared a UPS2 analysis dataset [PXD000331(Ahrne, et al., 2013)] which was generated by conducting three duplicate analyses of a UPS2 sample. MaxQuant is one of most widely used tools for label-free quantitative analyses. It uses a protein sequence DB search engine, called Andromeda. Therefore, analytical processes such as peptide and protein sequencing and quantification are conducted in a non-stop manner. By comparing the results of these two applications, we estimated the analytical performance of Epsilon-Q. MaxQuant identified a total of 32 protein pairs from the UPS2 dataset, with good matches to the expected

abundance ratios (Figure 6A, upper panel). Epsilon-Q identified 42 protein pairs; 10 more protein pairs than MaxQuant. Although there are some matches with expected ratios for low abundance proteins, most of them were not identified by MaxQuant (Figure 6B, lower panel). Both Epsilon-Q and MaxQuant show a good linearity (more than 0.99) for common 32 identified proteins with no significant difference. However, for replicated experiments of complex protein mixture, the difference between them was increased (see Figure 7). MaxQuant shows excellent matches with expected protein ratios but the number of quantifiable protein pairs seems declined (Figure 7, upper panel). Epsilon-Q shows more deviations of calculated protein abundance ratios than MaxQuant but it has much higher identified protein pairs (Figure 7, lower panel). These results demonstrate that Epsilon-Q exhibits results comparable with MaxQuant, but better detectability for low abundance proteins.

We also evaluated the quantitative reproducibility of Epsilon-Q using three replicated UPS2 samples that were spiked into *Drosophila* and *Mycoplasma* cell lysates [PXD000331(Ahrne, et al., 2013)]. The protein abundances were calculated by summing the peptide peak abundances assigned to each protein. The correlation of each replicative pair was estimated by linear regression. Figure 8 shows that the $R^2$ values for the pair correlations were between 0.97 and 0.99, indicating good quantitative reproducibility over a wide range of background complexities. Using equal amounts of the UPS1 and UPS2 samples independently spiked into *E.coli* cell lysate [PXD000602(Krey, *et al.*, 2014)], we estimated the quantitative performance of Epsilon-Q over a dynamic abundance range of proteins. The protein ratios were calculated using the UPS1 sample as a control. Figure 9 shows the log ratios calculated by Epsilon-Q against the log of the true ratios. The results indicate that ratios of low abundance proteins tend to be less accurate, but the

abundance ratios calculated by Epsilon-Q generally show good linearity with their true ratios.

## 3.6. Epsilon-Q Interface

Epsilon-Q system includes two sub-tools: “SimSPL Builder” and “Combo-Spec Search”. SimSPL Builder generates customized simSPL based on protein sequences. Users can generate simSPL step-by-step for use in Combo-Spec Search, or force one step to use a specific feature, such as merging two libraries or generating decoys (Figure 10A). Combo-Spec Search provides an interface to multiple spectral library searching (Figure 10B). All results are output to a text-based file (csv format).

# 4. Discussion

Spectral library searching has been shown to have better sensitivity than sequence DB searching methods. However, each spectral library has limited proteome coverage and using combined library increases the false-positive rate. Hence, we recommend that each spectral library search and FDR control should be performed independently. These limitations make hard to sequence whole proteins and the processing throughput of large data sets using spectral library search methods. Thus, Epsilon-Q system was designed to overcome some of these limitations. Using Epsilon-Q system, users can perform multiple spectral library searching. For the sequencing of novel and missing proteins, use of a customized simSPL can also improve searches of spectral libraries. Epsilon-Q provides not only the SimSPL Builder tool, which builds simSPLs, but also the Combo-Spec Search tool to support multiple spectral library searching.

In this study, we demonstrated that Epsilon-Q system exhibits greater detectability for peptides and proteins than other sequence DB-based searching tools. For those proteins identified, Epsilon-Q automatically calculates their abundance index based on the sum of their precursor ion intensities. Based on the maximum precursor peak and the peak-shape model (FWHM), Epsilon-Q detects a group of precursor peaks for each peptide and sums their intensities. Furthermore, Epsilon-Q shows good quantitative reproducibility and linearity performance for a variety of complex standard datasets. In conclusion, Epsilon-Q is an efficient tool for comparative proteome analysis based on multiple spectral libraries and label-free quantification. This software is implemented in

the C# language and is compatible with Windows operating systems with .NET framework 4.0 installed.

**Table 1.** List of datasets used in SUBJECT II study.

| Dataset | Instrument | Groups | Replicates | PXID (PXID) |
|---|---|---|---|---|
| UPS2 protein mix in different cell lysates | LTQ Orbitrap Velos | UPS2 only | Two replicates | PXD000331 *(Ahrne, et al., 2013)* |
| | | UPS2 (spiked in *Drosophila*) | Three replicates | |
| | | UPS2 (spiked in *Leptospira*) | Three replicates | |
| | | UPS2 (spiked in Mycoplasma) | Three replicates | |
| UPS1 and UPS2 protein mix in *E.coli* | LTQ Orbitrap Velos | UPS1 (spiked in *E.coli*) | Six fractions in each four replicates | PXD000602 (Krey, et al., 2014) |
| | | UPS2 (spiked in *E.coli*) | Six fractions in each four replicates | |
| UPS1 in Yeast | LTQ Orbitrap Velos | UPS1 in Yeast cell lysate | Three replicates for nine abundance group | PXD001819 (Ramus, et al., 2016) |

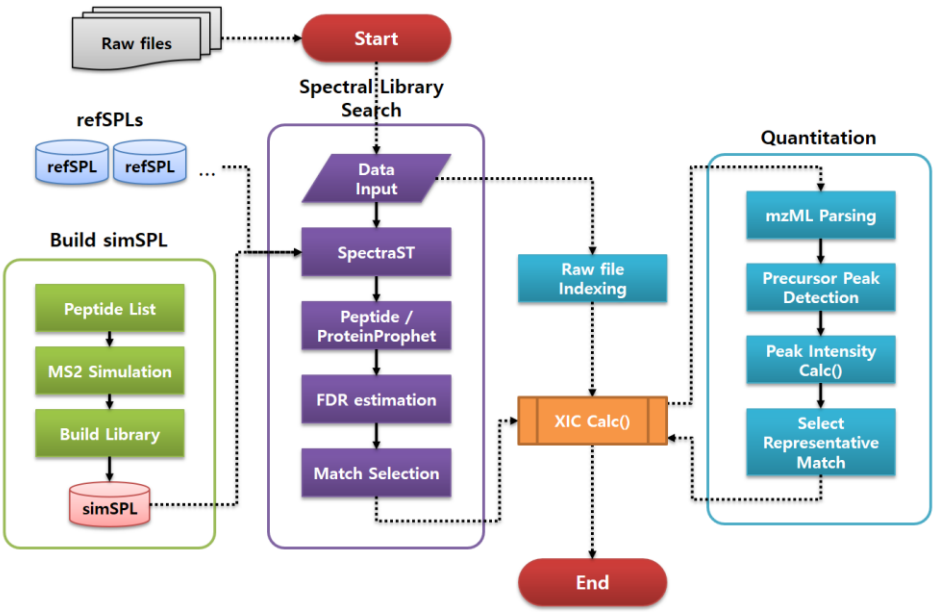



**Figure 1.** Epsilon-Q Workflow. Epsilon-Q supports SpectraST for spectral library searching. All processing steps, such as spectral library searching, statistical estimation, combining results, and protein abundance calculations, are automatically performed by Epsilon-Q.

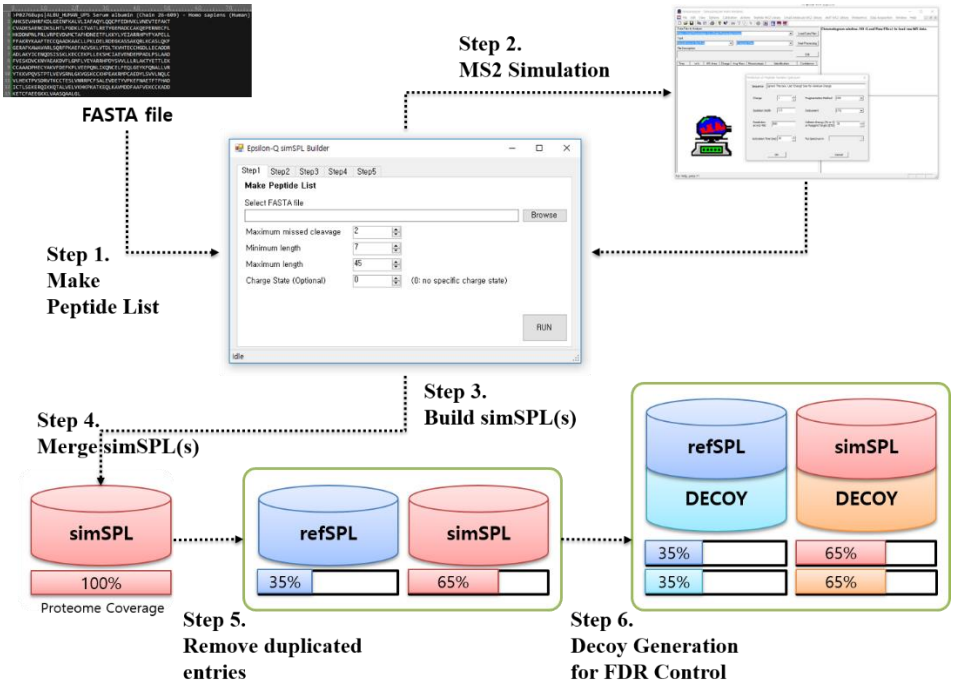



**Figure 2.** SimSPL Builder Workflow. The process is divided into six steps to generate a simSPL for use in the Combo-Spec Search method.

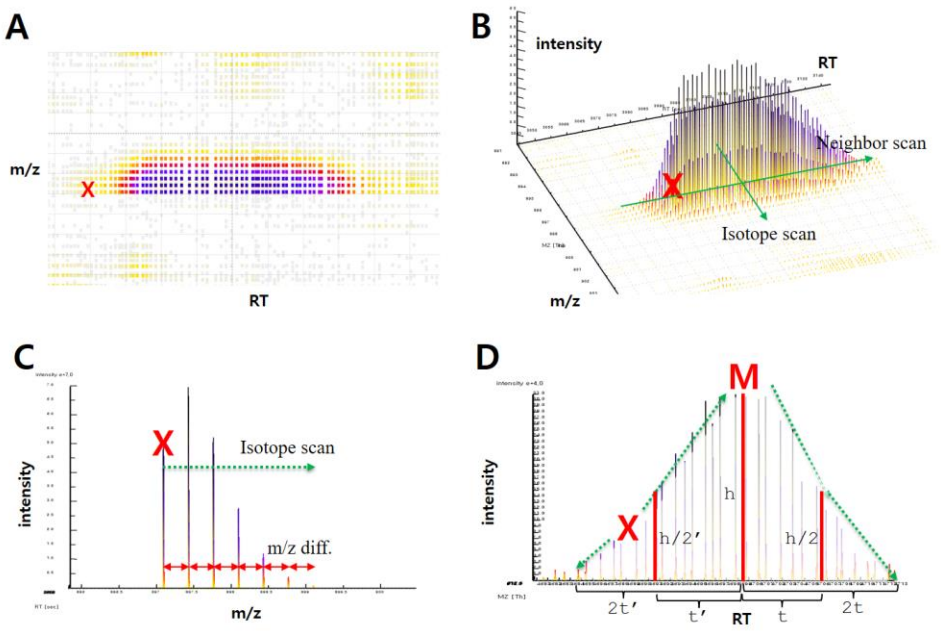



**Figure 3.** Precursor and Isotope Peak Detection. (A) Assigned precursor ion (red X) and group of neighbor peaks. (B) Neighbor peak and isotope peak scanning. (C) Isotope peak detection. (D) Determination of precursor peak range using the maximum peak and the width-at-half-height. Markers ‘X’ and ‘M’ represent identified precursor ion peak and maximum ion peak, respectively.

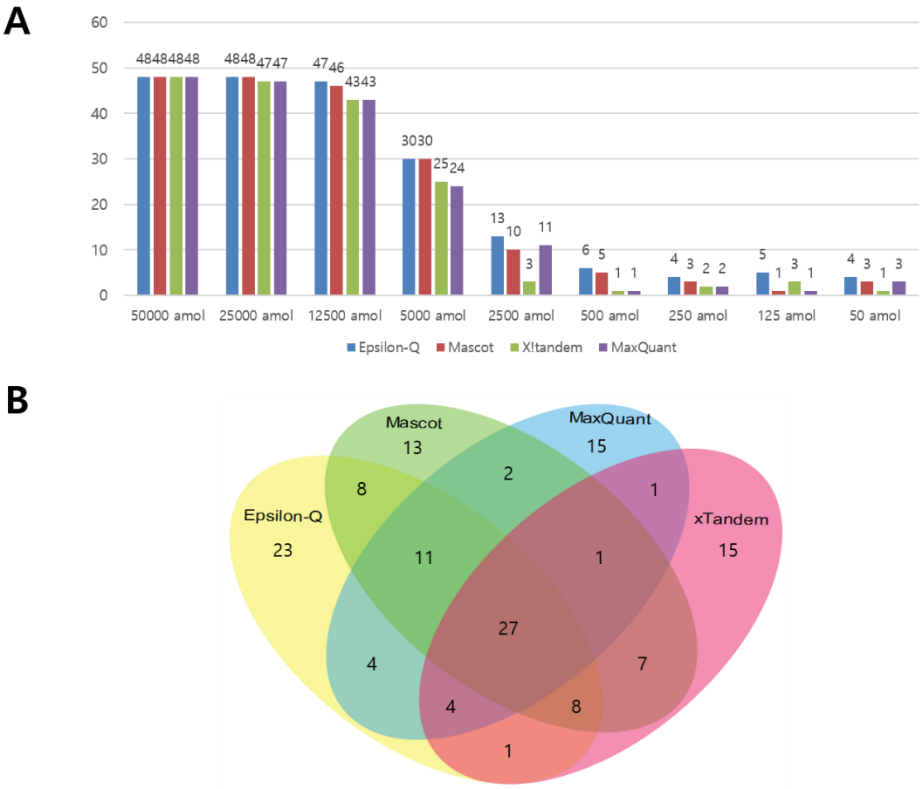



**Figure 4.** Number of proteins and distinct peptide sequences identified by different analytical tools. (A) Number of proteins identified by each tool at different concentrations. (B) Four-way Venn diagram showing the overlap between four tandem MS search engines.

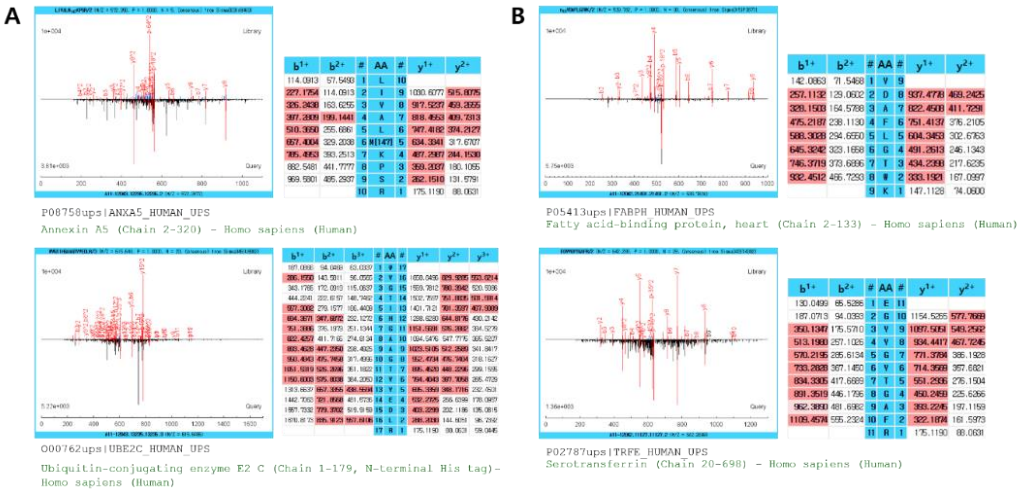



**Figure 5.** Examples of spectrum-to-spectrum matches detected by Epsilon-Q. (A) Spectrum-to-spectrum matches using the NIST UPS spectral library. (B) Spectrum-to-spectrum matches using a simSPL.

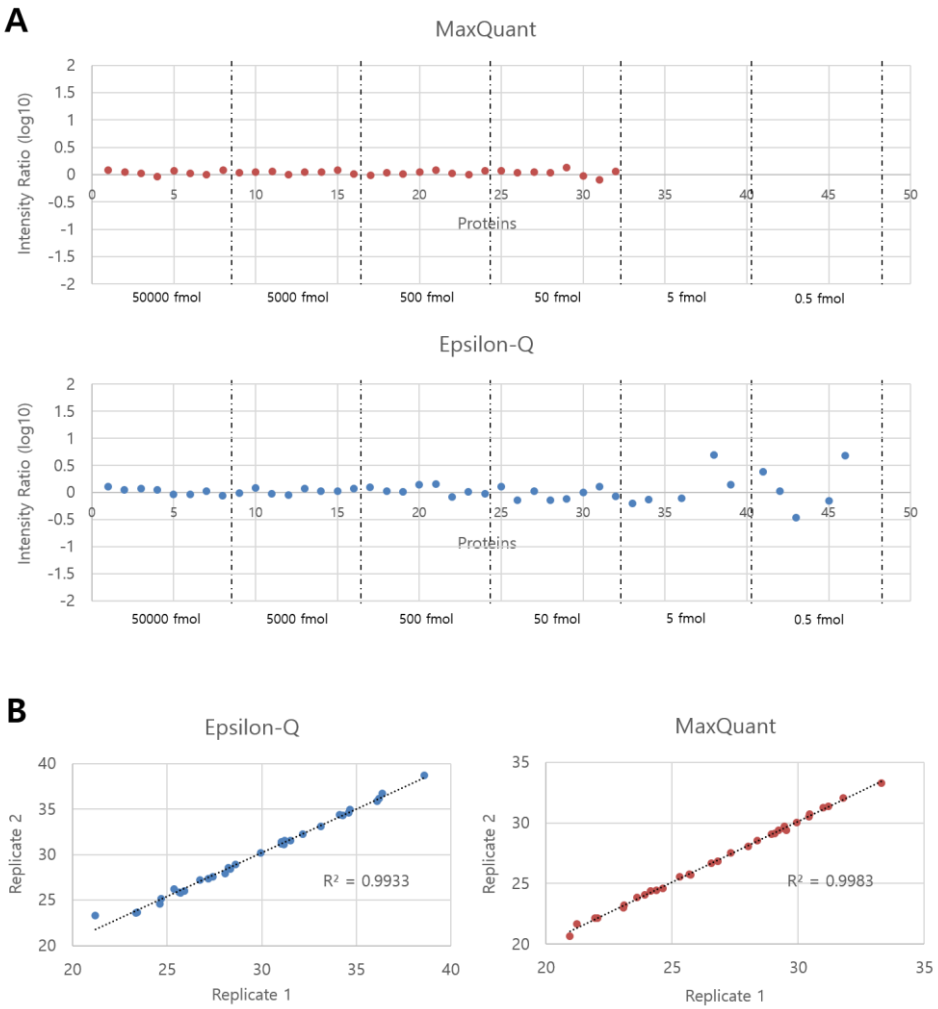



**Figure 6.** Comparison of the quantitative analytical performances of Epsilon-Q and MaxQuant. Two replicate experiment data sets of UPS2 were used in this test and expected ratio is 0. (A) Shown here is abundance ratios of 48 UPS proteins by Epsilon-Q and MaxQuant. (B) Quantitative correlation of 32 common identified proteins by Epsilon-Q and MaxQuant.

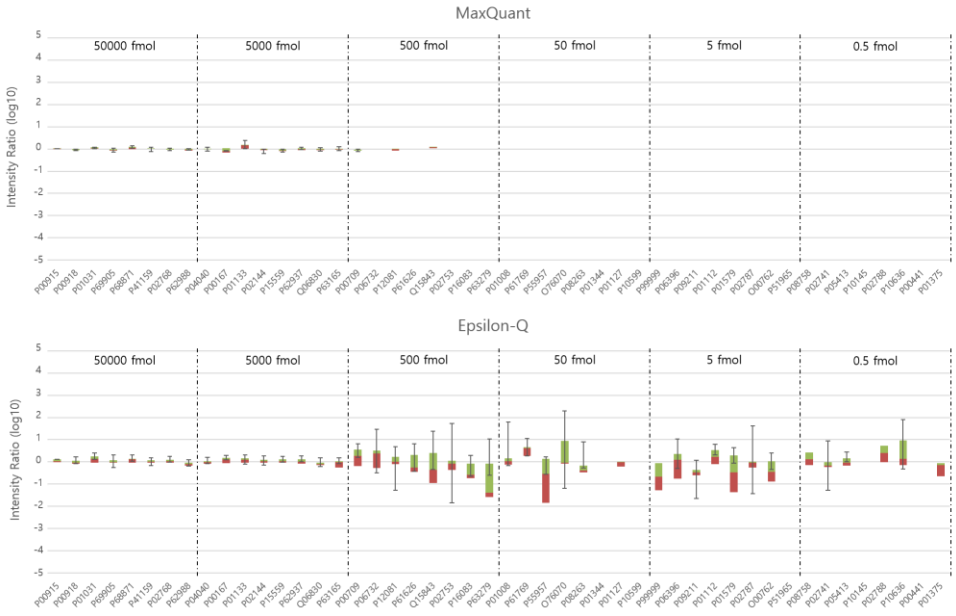



**Figure 7.** Comparison of quantitative analytical performance for complex MS data sets between Epsilon-Q and MaxQuant. UPS2 sample spiked into *Drosophila* cell lysates and UPS2 spiked into *Mycoplasma* cell lysates were used in this test (Ahrné, E *et al*., PXD000331(Ahrne, et al., 2013)). Each data set contains three replicate samples. Forty eight UPS protein abundance ratios were calculated from total 15 pairs of 6 experiments by Epsilon-Q and MaxQuant. The expected ratio is 0.

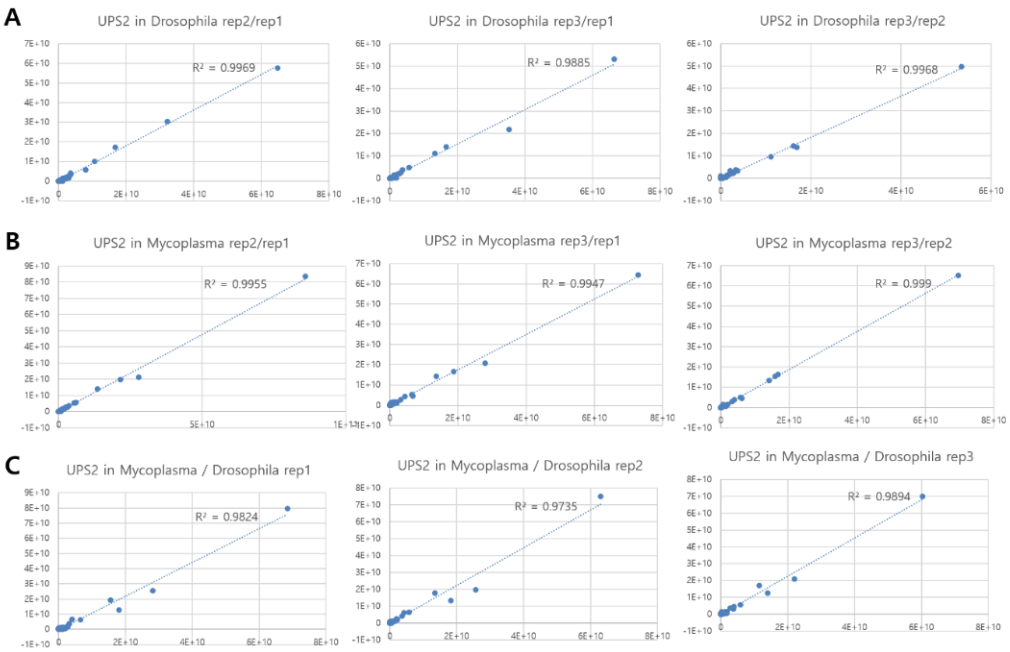



**Figure 8.** Scatter plots of replicative experimental pairs to evaluate analytical reproducibility. (A) Replicative experimental pairs in UPS2 samples spiked into *Drosophila* lysates. (B) Replicative experimental pairs in UPS2 samples spiked into *Mycoplasma* lysates. (C) Replicative experiment pairs in UPS2 samples spiked into each *Drosophila* and *Mycoplasma* lysates.

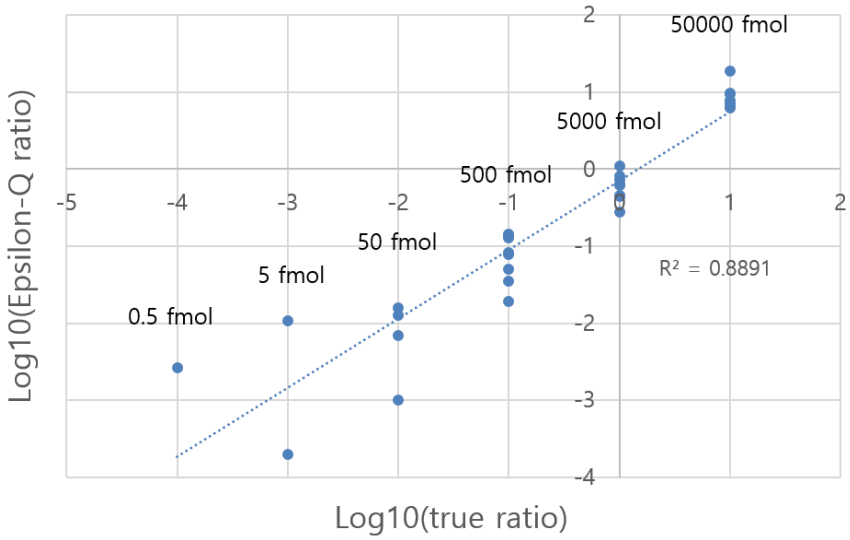



**Figure 9.** Scatter plot of log scale intensity ratios of UPS2 samples versus UPS1 samples spiked in *E.coli* determined by Epsilon-Q.

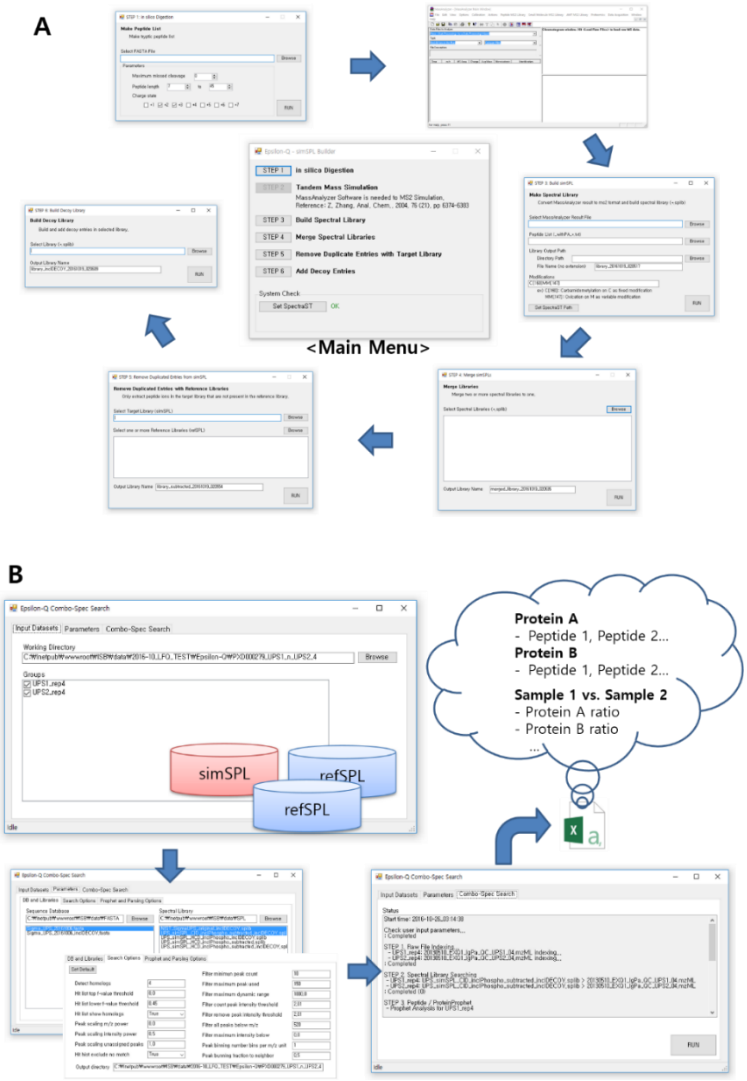



**Figure 10.** Epsilon-Q Interface. (A) SimSPL Builder Interface. This tool provides six steps to build a simSPL. (B) Combo-Spec Search Interface. Here, users can select search engines and peptide spectral libraries. Based on the input parameters, Epsilon-Q automatically performs multiple library searching, statistical estimation and results parsing.

# Conclusions

Although the rigorous protein search analyzes were carried out on MS data produced under the instrument's optimal performance conditions, it is inevitable that some proteins will remain undetected. It is why we need to develop a better search strategy and analytical software that provides greater sensitivity and more accurate analysis in the search for missing proteins. Spectral library searching shows better sensitivity than sequence DB searching method. However, each spectral library has limited proteome coverage (30-40%) and some false-positive rate. We suggest that combination of multiple spectral libraries with different proteome coverage called Combo-Spec Search method could be one solution to avoid the limitation. This study demonstrates that the application of Combo-Spec Search method to a previously analyzed dataset (Lee, et al., 2013) can present additional opportunities to identify missing proteins that have never been detected by sequence DB searches.

We also develop the new analytical software called Epsilon-Q. Using the Epsilon-Q, users can perform multiple spectral library searching. Especially, for novel and missing proteins sequencing, using customized simSPL could be a complement to further remedy imperfections of spectral libraries. Epsilon-Q provides SimSPL Builder which builds simSPL. Epsilon-Q also provide Combo-Spec Search tool to support multiple spectral library searching. Here we demonstrated that Epsilon-Q shows more improved detectability for peptides and proteins that other sequence DB-based searching tools. For identified proteins, Epsilon-Q automatically calculates its abundance index based on the sum of precursor ion intensities. In this study, Epsilon-Q shows good quantitative

reproducibility and linearity performance for variety complex of standard datasets. In summary, Epsilon-Q is an efficient tool for comparative proteome analysis based on multiple spectral libraries and label-free quantification. This software is executable on Windows operating system.

# References

Aebersold, R.; Mann, M., 2003, Mass spectrometry-based proteomics, 422(6928): 198-207

Ahrne, E.; Molzahn, L.; Glatter, T.; Schmidt, A., 2013, Critical assessment of proteome-wide label-free absolute abundance estimation strategies, 13(17): 2567-78

Bantscheff, M.; Schirle, M.; Sweetman, G.; Rick, J.; Kuster, B., 2007, Quantitative mass spectrometry in proteomics: a critical review, 389(4): 1017-31

Chait, B. T., 2006, Chemistry. Mass spectrometry: bottom-up or top-down?, 314(5796): 65-6

Chambers, M. C.; Maclean, B.; Burke, R.; Amodei, D.; Ruderman, D. L.; Neumann, S.; Gatto, L.; Fischer, B.; Pratt, B.; Egertson, J.; Hoff, K.; Kessner, D.; Tasman, N.; Shulman, N.; Frewen, B.; Baker, T. A.; Brusniak, M. Y.; Paulse, C.; Creasy, D.; Flashner, L.; Kani, K.; Moulding, C.; Seymour, S. L.; Nuwaysir, L. M.; Lefebvre, B.; Kuhlmann, F.; Roark, J.; Rainer, P.; Detlev, S.; Hemenway, T.; Huhmer, A.; Langridge, J.; Connolly, B.; Chadick, T.; Holly, K.; Eckels, J.; Deutsch, E. W.; Moritz, R. L.; Katz, J. E.; Agus, D. B.; MacCoss, M.; Tabb, D. L.; Mallick, P., 2012, A cross-platform toolkit for mass spectrometry and proteomics, 30(10): 918-20

Chelius, D.; Bondarenko, P. V., 2002, Quantitative profiling of proteins in complex mixtures using liquid chromatography and mass spectrometry, 1(4): 317-23

Cho, J. Y.; Lee, H. J.; Jeong, S. K.; Kim, K. Y.; Kwon, K. H.; Yoo, J. S.; Omenn, G. S.; Baker, M. S.; Hancock, W. S.; Paik, Y. K., 2015, Combination of Multiple Spectral Libraries Improves the Current Search Methods Used to Identify Missing Proteins in the Chromosome-Centric Human Proteome Project, 14(12): 4959-66

Cox, J.; Mann, M., 2008, MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification, 26(12): 1367-72

Craig, R.; Beavis, R. C., 2004, TANDEM: matching proteins with tandem mass spectra, 20(9): 1466-7

Craig, R.; Cortens, J. C.; Fenyo, D.; Beavis, R. C., 2006, Using annotated peptide mass spectrum libraries for protein identification, 5(8): 1843-9

Desiere, F.; Deutsch, E. W.; Nesvizhskii, A. I.; Mallick, P.; King, N. L.; Eng, J. K.; Aderem, A.; Boyle, R.; Brunner, E.; Donohoe, S.; Fausto, N.; Hafen, E.; Hood, L.; Katze, M. G.; Kennedy, K. A.; Kregenow, F.; Lee, H.; Lin, B.; Martin, D.; Ranish, J. A.; Rawlings, D. J.; Samelson, L. E.; Shiio, Y.; Watts, J. D.; Wollscheid, B.; Wright, M. E.; Yan, W.; Yang, L.; Yi, E. C.; Zhang, H.; Aebersold, R., 2005, Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry, 6(1): R9

Deutsch, E. W.; Mendoza, L.; Shteynberg, D.; Slagel, J.; Sun, Z.; Moritz, R. L., 2015a, Trans-Proteomic Pipeline, a standardized data processing pipeline for large-scale reproducible proteomics informatics, 9(7-8): 745-54

Deutsch, E. W.; Sun, Z.; Campbell, D.; Kusebauch, U.; Chu, C. S.; Mendoza, L.; Shteynberg, D.; Omenn, G. S.; Moritz, R. L., 2015b, State of the Human Proteome in 2014/2015 As Viewed through PeptideAtlas: Enhancing Accuracy and Coverage through the AtlasProphet, 14(9): 3461-73

Dhingra, V.; Gupta, M.; Andacht, T.; Fu, Z. F., 2005, New frontiers in proteomics research: a perspective, 299(1-2): 1-18

Eng, J. K.; Jahan, T. A.; Hoopmann, M. R., 2013, Comet: an open-source MS/MS sequence database search tool, 13(1): 22-4

Eng, J. K.; McCormack, A. L.; Yates, J. R., 1994, An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database, 5(11): 976-89

Fenyo, D.; Beavis, R. C., 2003, A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes, 75(4): 768-74

Frewen, B. E.; Merrihew, G. E.; Wu, C. C.; Noble, W. S.; MacCoss, M. J., 2006, Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries, 78(16): 5678-84

Fu, X.; Gharib, S. A.; Green, P. S.; Aitken, M. L.; Frazer, D. A.; Park, D. R.; Vaisar, T.; Heinecke, J. W., 2008, Spectral index for assessment of differential protein expression in shotgun proteomics, 7(3): 845-54

Geer, L. Y.; Markey, S. P.; Kowalak, J. A.; Wagner, L.; Xu, M.; Maynard, D. M.; Yang, X.;

Shi, W.; Bryant, S. H., 2004, Open mass spectrometry search algorithm, 3(5): 958-64

Gygi, S. P.; Rochon, Y.; Franza, B. R.; Aebersold, R., 1999, Correlation between protein and mRNA abundance in yeast, 19(3): 1720-30

Hu, Y.; Lam, H., 2013, Expanding tandem mass spectral libraries of phosphorylated peptides: advances and applications, 12(12): 5971-7

Hu, Y.; Li, Y.; Lam, H., 2011, A semi-empirical approach for predicting unobserved peptide MS/MS spectra from spectral libraries, 11(24): 4702-11

Ishihama, Y.; Oda, Y.; Tabata, T.; Sato, T.; Nagasu, T.; Rappsilber, J.; Mann, M., 2005, Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein, 4(9): 1265-72

James, P., 1997, Protein identification in the post-genome era: the rapid rise of proteomics, 30(4): 279-331

Ji, C.; Arnold, R. J.; Sokoloski, K. J.; Hardy, R. W.; Tang, H.; Radivojac, P., 2013, Extending the coverage of spectral libraries: a neighbor-based approach to predicting intensities of peptide fragmentation spectra, 13(5): 756-65

Junker, J.; Bielow, C.; Bertsch, A.; Sturm, M.; Reinert, K.; Kohlbacher, O., 2012, TOPPAS: a graphical workflow editor for the analysis of high-throughput proteomics data, 11(7): 3914-20

Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R., 2002, Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search, 74(20): 5383-92

Kim, S.; Pevzner, P. A., 2014, MS-GF+ makes progress towards a universal database search tool for proteomics, 5(5277

Krey, J. F.; Wilmarth, P. A.; Shin, J. B.; Klimek, J.; Sherman, N. E.; Jeffery, E. D.; Choi, D.; David, L. L.; Barr-Gillespie, P. G., 2014, Accurate label-free protein quantitation with high- and low-resolution mass spectrometers, 13(2): 1034-44

Lam, H., 2011, Building and searching tandem mass spectral libraries for peptide identification, 10(12): R111 008565

Lam, H.; Aebersold, R., 2011, Building and searching tandem mass (MS/MS) spectral libraries for peptide identification in proteomics, 54(4): 424-31

Lam, H.; Deutsch, E. W.; Aebersold, R., 2010, Artificial decoy spectral libraries for false discovery rate estimation in spectral library searching in proteomics, 9(1): 605-10

Lam, H.; Deutsch, E. W.; Eddes, J. S.; Eng, J. K.; King, N.; Stein, S. E.; Aebersold, R., 2007, Development and validation of a spectral library searching method for peptide identification from MS/MS, 7(5): 655-67

Lam, H.; Deutsch, E. W.; Eddes, J. S.; Eng, J. K.; Stein, S. E.; Aebersold, R., 2008, Building consensus spectral libraries for peptide identification in proteomics, 5(10): 873-5

Lane, L.; Bairoch, A.; Beavis, R. C.; Deutsch, E. W.; Gaudet, P.; Lundberg, E.; Omenn, G. S., 2014, Metrics for the Human Proteome Project 2013-2014 and strategies for finding missing proteins, 13(1): 15-20

Lee, H. J.; Jeong, S. K.; Na, K.; Lee, M. J.; Lee, S. H.; Lim, J. S.; Cha, H. J.; Cho, J. Y.; Kwon, J. Y.; Kim, H.; Song, S. Y.; Yoo, J. S.; Park, Y. M.; Kim, H.; Hancock, W. S.; Paik, Y. K., 2013, Comprehensive genome-wide proteomic analysis of human placental tissue for the Chromosome-Centric Human Proteome Project, 12(6): 2458-66

Lill, J., 2003, Proteomic tools for quantitation by mass spectrometry, 22(3): 182-94

Liu, H.; Sadygov, R. G.; Yates, J. R., 3rd, 2004, A model for random sampling and estimation of relative protein abundance in shotgun proteomics, 76(14): 4193-201

Maher, B., 2012, ENCODE: The human encyclopaedia, 489(7414): 46-8

Nesvizhskii, A. I., 2014, Proteogenomics: concepts, applications and computational strategies, 11(11): 1114-25

Old, W. M.; Meyer-Arendt, K.; Aveline-Wolf, L.; Pierce, K. G.; Mendoza, A.; Sevinsky, J. R.; Resing, K. A.; Ahn, N. G., 2005, Comparison of label-free methods for quantifying human proteins by shotgun proteomics, 4(10): 1487-502

Paik, Y. K.; Jeong, S. K.; Omenn, G. S.; Uhlen, M.; Hanash, S.; Cho, S. Y.; Lee, H. J.; Na, K.; Choi, E. Y.; Yan, F.; Zhang, F.; Zhang, Y.; Snyder, M.; Cheng, Y.; Chen, R.; Marko-Varga, G.; Deutsch, E. W.; Kim, H.; Kwon, J. Y.; Aebersold, R.; Bairoch, A.; Taylor, A. D.;

Kim, K. Y.; Lee, E. Y.; Hochstrasser, D.; Legrain, P.; Hancock, W. S., 2012a, The Chromosome-Centric Human Proteome Project for cataloging proteins encoded in the genome, 30(3): 221-3

Paik, Y. K.; Omenn, G. S.; Uhlen, M.; Hanash, S.; Marko-Varga, G.; Aebersold, R.; Bairoch, A.; Yamamoto, T.; Legrain, P.; Lee, H. J.; Na, K.; Jeong, S. K.; He, F.; Binz, P. A.; Nishimura, T.; Keown, P.; Baker, M. S.; Yoo, J. S.; Garin, J.; Archakov, A.; Bergeron, J.; Salekdeh, G. H.; Hancock, W. S., 2012b, Standard guidelines for the chromosome-centric human proteome project, 11(4): 2005-13

Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S., 1999, Probability-based protein identification by searching sequence databases using mass spectrometry data, 20(18): 3551-67

Ramus, C.; Hovasse, A.; Marcellin, M.; Hesse, A. M.; Mouton-Barbosa, E.; Bouyssie, D.; Vaca, S.; Carapito, C.; Chaoui, K.; Bruley, C.; Garin, J.; Cianferani, S.; Ferro, M.; Dorssaeler, A. V.; Burlet-Schiltz, O.; Schaeffer, C.; Coute, Y.; Gonzalez de Peredo, A., 2016, Spiked proteomic standard dataset for testing label-free quantitative software and statistical methods, 6(286-94

Rousseeuw, P. J.; Croux, C., 1993, Alternatives to the Median Absolute Deviation, 88(424): 1273-1283

Shteynberg, D.; Deutsch, E. W.; Lam, H.; Eng, J. K.; Sun, Z.; Tasman, N.; Mendoza, L.; Moritz, R. L.; Aebersold, R.; Nesvizhskii, A. I., 2011, iProphet: multi-level integrative

analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates, 10(12): M111 007690

Shteynberg, D.; Nesvizhskii, A. I.; Moritz, R. L.; Deutsch, E. W., 2013, Combining results of multiple search engines in proteomics, 12(9): 2383-93

Steen, H.; Mann, M., 2004, The ABC's (and XYZ's) of peptide sequencing, 5(9): 699-711

Stein, S. E.; Scott, D. R., 1994, Optimization and testing of mass spectral library search algorithms for compound identification, 5(9): 859-66

Suni, V.; Imanishi, S. Y.; Maiolica, A.; Aebersold, R.; Corthals, G. L., 2015a, Confident Site Localization Using a Simulated Phosphopeptide Spectral Library,

Suni, V.; Imanishi, S. Y.; Maiolica, A.; Aebersold, R.; Corthals, G. L., 2015b, Confident site localization using a simulated phosphopeptide spectral library, 14(5): 2348-59

Tabb, D. L.; Fernando, C. G.; Chambers, M. C., 2007, MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis, 6(2): 654-61

Wang, W.; Zhou, H.; Lin, H.; Roy, S.; Shaler, T. A.; Hill, L. R.; Norton, S.; Kumar, P.; Anderle, M.; Becker, C. H., 2003, Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards, 75(18): 4818-26

Washburn, M. P.; Wolters, D.; Yates, J. R., 3rd, 2001, Large-scale analysis of the yeast proteome by multidimensional protein identification technology, 19(3): 242-7

Wolters, D. A.; Washburn, M. P.; Yates, J. R., 3rd, 2001, An automated multidimensional protein identification technology for shotgun proteomics, 73(23): 5683-90

Yates, J. R., 3rd; Morgan, S. F.; Gatlin, C. L.; Griffin, P. R.; Eng, J. K., 1998, Method to compare collision-induced dissociation spectra of peptides: potential for library searching and subtractive analysis, 70(17): 3557-65

Yen, C. Y.; Houel, S.; Ahn, N. G.; Old, W. M., 2011, Spectrum-to-spectrum searching using a proteome-wide spectral library, 10(7): M111 007666

Yen, C. Y.; Meyer-Arendt, K.; Eichelberger, B.; Sun, S.; Houel, S.; Old, W. M.; Knight, R.; Ahn, N. G.; Hunter, L. E.; Resing, K. A., 2009, A simulated MS/MS library for spectrum-to-spectrum searching in large scale identification of proteins, 8(4): 857-69

Zhang, X.; Li, Y.; Shao, W.; Lam, H., 2011, Understanding the improved sensitivity of spectral library searching over sequence database searching in proteomics data analysis, 11(6): 1075-85

Zhang, Z., 2004, Prediction of low-energy collision-induced dissociation spectra of peptides, 76(14): 3908-22

Zhang, Z., 2005, Prediction of low-energy collision-induced dissociation spectra of peptides with three or more charges, 77(19): 6364-73

Zhang, Z.; Wu, S.; Stenoien, D. L.; Pasa-Tolic, L., 2014, High-throughput proteomics, 7: 427-54

## Abstract in Korean

# 질량 분석 데이터를 활용한 단백질의 정성 및 정량 분석용 생물정보학 플랫폼의 개발

**Jin-Young Cho**

**Department of Integrated OMICS for Biomedical Science of World Class University**

**The Graduate School**

**Yonsei University**

인간에게는 약 29 억 염기 쌍의 길이의 유전체 서열이 있으며, 여기에 약 20,000 여 개의 대표 단백질들의 발현 정보가 들어 있다고 알려져 있다. 하지만, 이 가운데 약 15% 정도에 해당하는 단백질들은 실험적인 존재 규명 근거가 미비하여 “미확인 단백질 (missing protein)”이라 불린다. 이 단백질들은 극히 국소적인 부분에서 미량으로 발현되는 이유로 발견이 어려울 것이라 추정되며, 단백질 분석의 기술적인 한계도 여기에 일조한다. 이를테면 현재의 질량분석 기법은 복잡한 생물학적인 시료 분석을 완벽하게 분석하는데 한계가 있으며, 서열 DB 검색의 정확도와 펩타이드 라이브러리의

제한된 단백질 분석 가능 범위도 해결해야 할 과제이다. 이러한 문제를 해결하기 위해 우리는 기존 spectral library (refSPL) 의 통합 (iRefSPL) 과 가상의 spectral library (simSPL) 제작을 통해 기존의 spectral library 분석의 한계를 극복하는 “Combo-Spec Search” 기법을 고안하였다. 이 기법은 기존의 spectral library 에 포함되지 않은 펩타이드 서열들의 가상 질량분석 스펙트럼 정보를 포함한 simSPL 을 활용함으로써 spectral library 의 분석 가능한 단백질 범위를 극대화 한다. 우리는 iRefSPL 과 simSPL 의 병행 사용이 상보적인 관계로 작용하여 더 많은 펩타이드 및 단백질을 탐지하는데 도움이 됨을 확인하였으며, 과거 단백질 서열 DB 검색으로 분석한 바 있는 인간 태반의 단백질 질량분석 데이터를 Combo-Spec Search 로 재분석하여 12 개의 미확인 단백질들에 대한 단서를 추가적으로 확보할 수 있었다.

우리는 Combo-Spec Search 기법을 소프트웨어적으로 자동화한 “Epsilon-Q” 소프트웨어를 개발하였다. 기존의 복잡하고 번거로운 다중 펩타이드 라이브러리 검색과 데이터 통합 과정을 이 소프트웨어를 통해 자동으로 수행할 수 있다. 이 프로그램은 표준 질량분석 데이터 포멧을 지원하여 보편적인 사용자 시스템 환경에서 사용이 가능하며, 자동으로 파일 인덱싱, 복수의 펩타이드 라이브러리 검색, 분석 결과 통합 기능을 제공할 뿐만 아니라 개별 단백질들의 양적인 발현 비교를 위한 정량지표 계산도 가능하다. 그래픽 유저 인터페이스를 제공하는 이 프로그램은 단백질의 정성 및 정량분석에 탁월한 성능을 보여줌을 확인하였으며, 이러한 결과를 근거로

우리는 Combo-Spec Search 기법과 이에 기반한 Epsilon-Q 가 복잡한 생물학적 시료로부터 미량의 단백질을 탐지하고 비교 정량을 수행하는데 효과적인 수단이 될 수 있으리라 기대한다.

핵심 되는 말: 상향식 단백질 분석 접근법, Chromosome-centric Human Proteome Project, Combo-Spec Search method, Epsilon-Q, 비표지 정량분석 기법, 질량분석법, 미확인 단백질, 단백질 동정, 단백체학, 서열 데이터베이스 검색법, 펩타이드 라이브러리