Data Quality Control in Peptide Identification

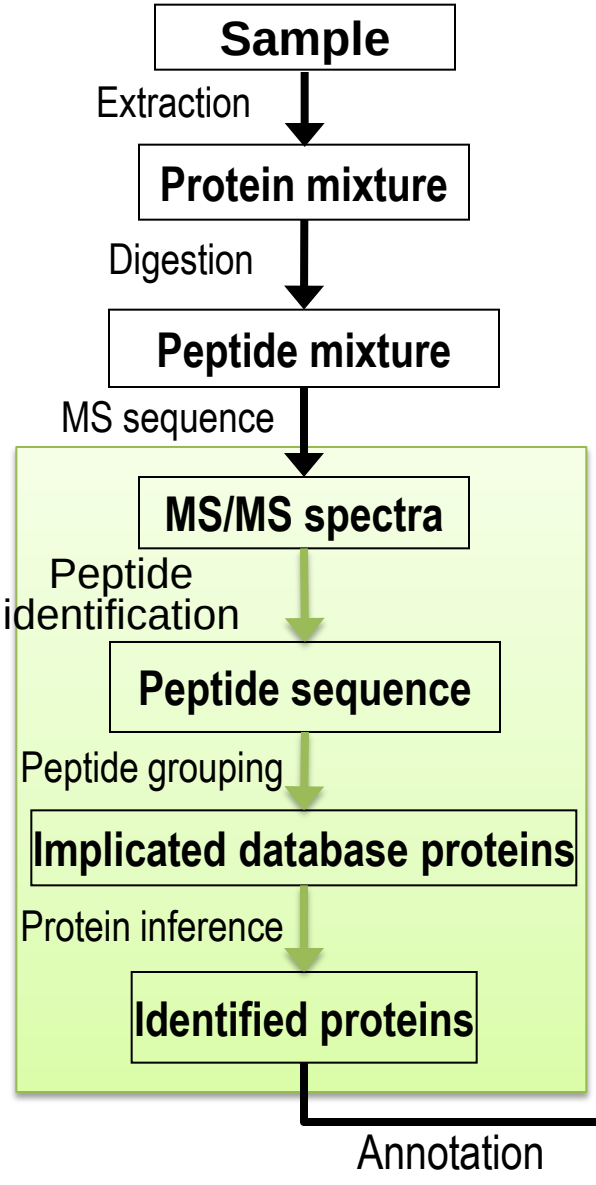
Yunping Zhu
State Key Laboratory of Proteomics
Beijing Proteome Research Center
Beijing Institute of Radiation Medicine

Beijing, 2010-11-11

- Data quality control for peptide identification in shotgun proteomics
- Evaluation of the effects of decoy design, search strategy, and mass tolerance on the accuracy and sensitivity of peptide identifications in shotgun proteomics

Data quality control for peptide identification in shotgun proteomics

Shotgun proteomics for peptide and protein identification



Analyze complex protein mixtures using shotgun proteomics

• Simplify MS/MS sequencing • Simplify sample handling and increase the overall data throughput	• Lost connectivity between peptides and proteins • Complicate computational analysis and biological interpretation

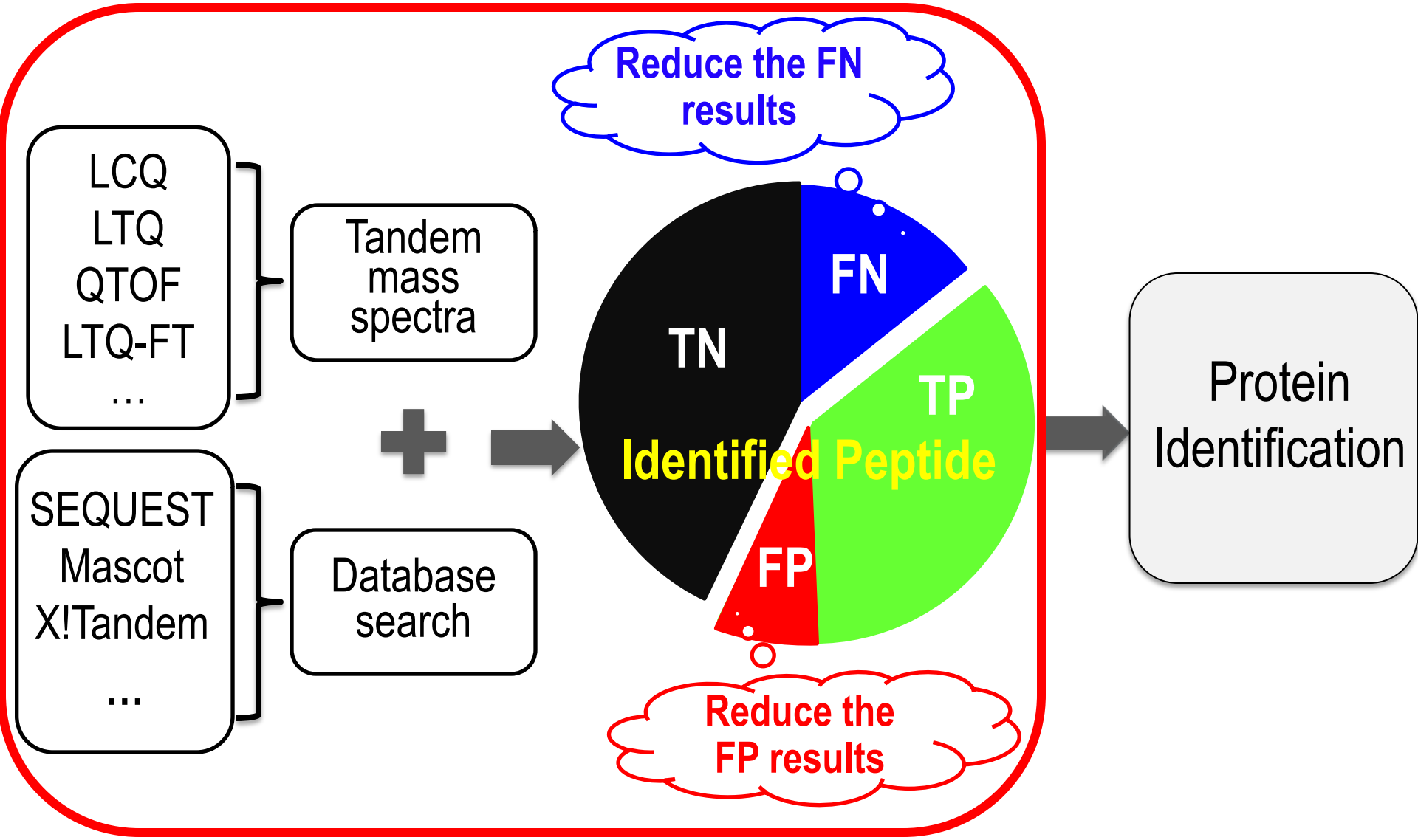
Quality control in data analysis process

Mol Cell Proteomics. 2005, 4(10):1419-40

Our focus

- Our group has been systematically studying the validation of database search results identified by shotgun proteomics.
  - A new strategy to filter out false positive identifications of peptides in SEQUEST database search results. *Proteomics*. 2007 19;7(22):4036-44.
  - A nonparametric model for quality control of database search results in shotgun proteomics. *BMC Bioinformatics* 2008, 9:29.
  - Mass measurement errors of Fourier-transform mass spectrometry (FTMS): distribution, recalibration, and application. *J Proteome Res.* 2009 Feb;8(2):849-59.
  - Bayesian nonparametric model for the validation of peptide identification in shotgun proteomics. *Mol. Cell. proteomics.* 2009, 8(3): 547-57.
  - Combination of new features improves peptide identification by Mascot in shotgun proteomics. (*Proteomics*, accepted)
- Make use of available features which were typically ignored could benefit data analysis process.
- Combination of new features with an appropriate framework can improve the sensitivity of the filtration methods.

Quality control of peptide identification in shotgun proteomics



TP, true positive; TN, true negative; FP, false positive; FN, false negative.

Mass measurement errors of Fourier-transform mass spectrometry (FTMS): distribution, recalibration, and application

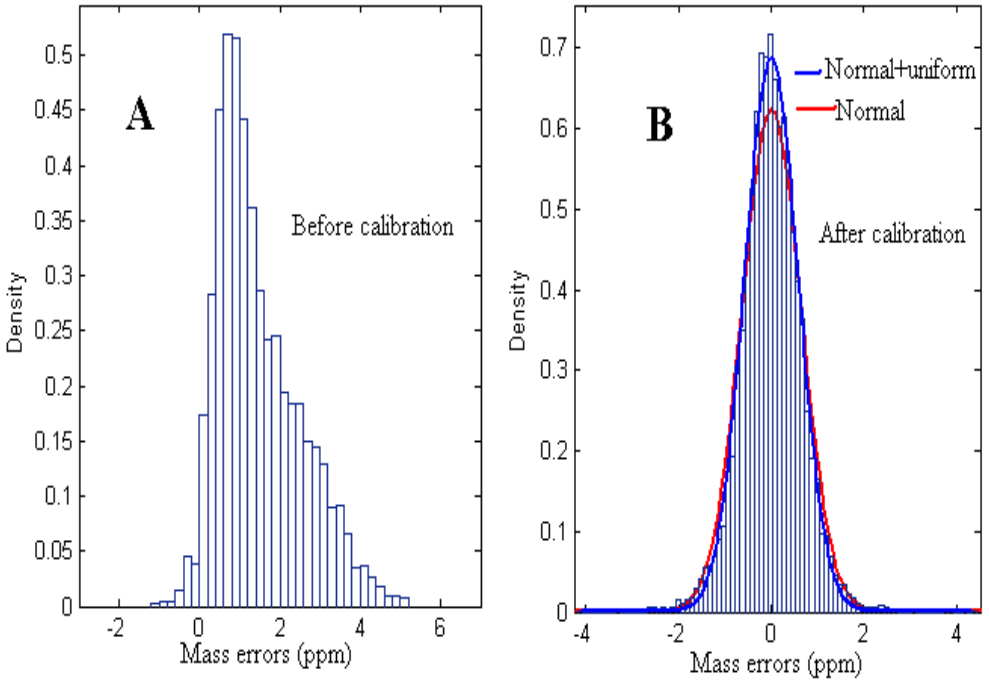
- Conducted a comprehensive investigation of the distribution of precursor ion mass error for the LTQ-FT platform;
- Developed an automatic GUI software tool, FTDR, for the recalibration of LTQ-FT MS data;
- Proposed and applied a new strategy LDSF to recalibrate the MS/MS data and improve peptide identification.

Improve the mass recalibration of FTMS data

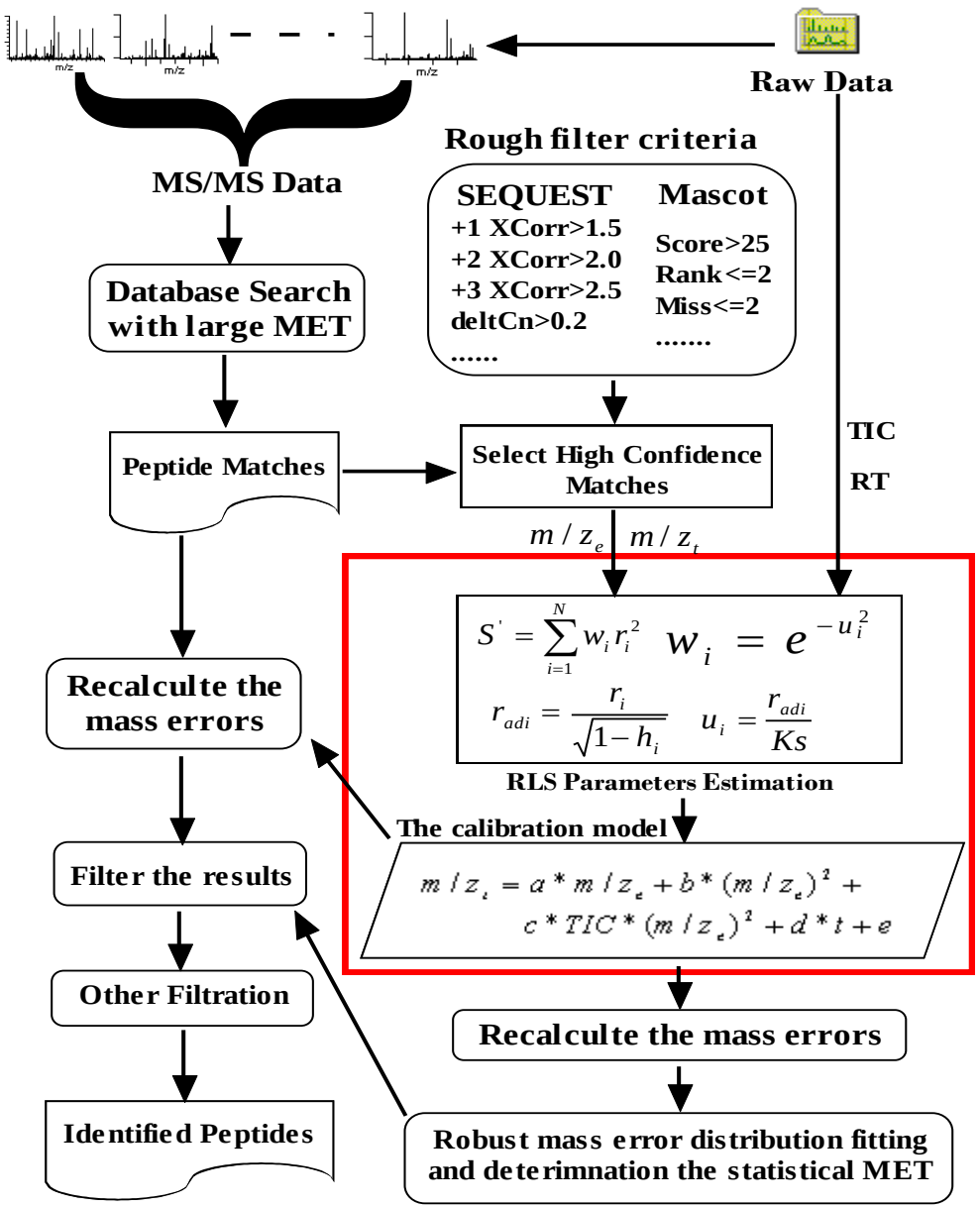
- An improved recalibration formula:

$$m/z_t = a/f + b/f^2 + c*TIC/f^2 + d*t + e$$

- **FTDR**，Fourier-transform data recalibration
  - An automated, multi-thread program
  - SEQUEST*.out and Mascot *.htm
  - Thermo *.RAW format, mzXML, or mzData

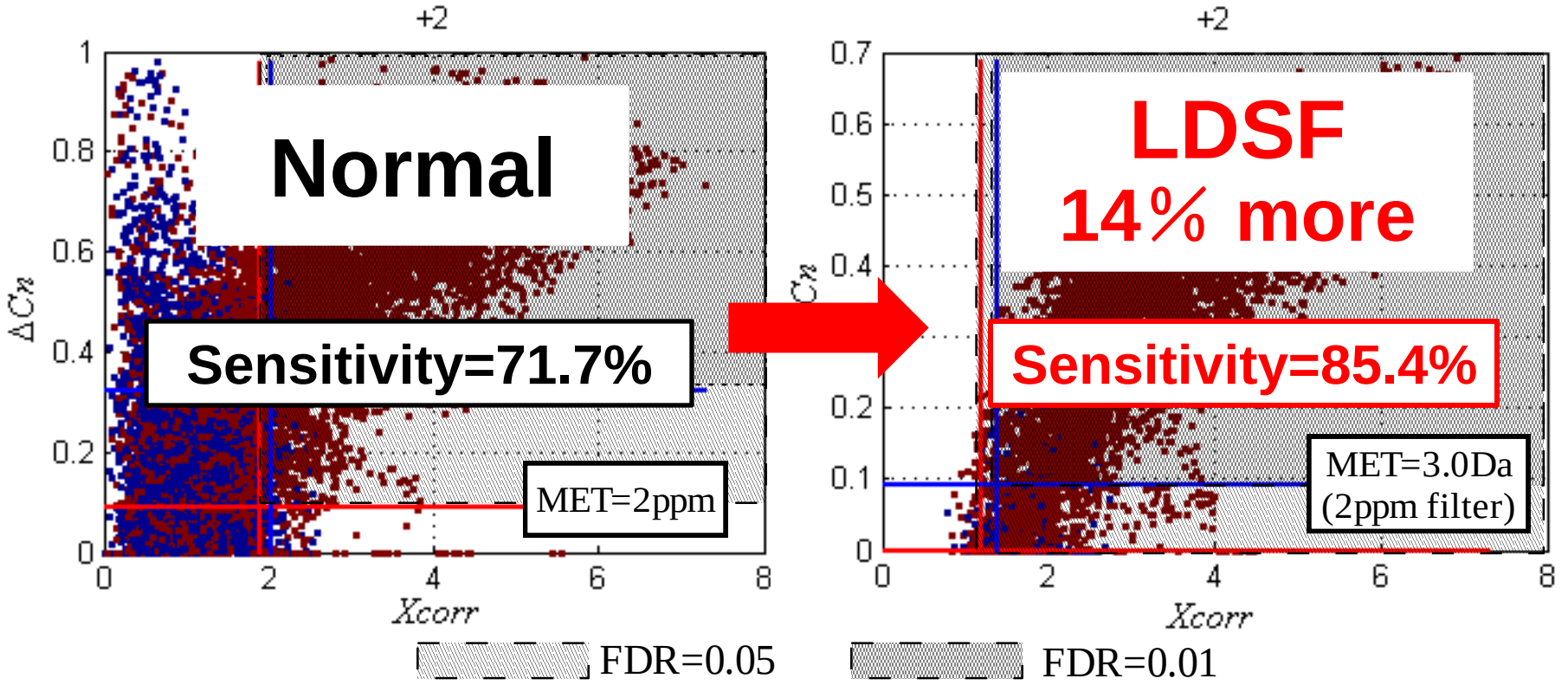


LDSF - Large MET database search followed by small MET filtration



- A new strategy to determinate the database search MET and validate the database search results
  - A large MET database search
  - Estimate the statistical MET
  - Recalibrate mass errors
  - Filter all the database search results with the statistical MET

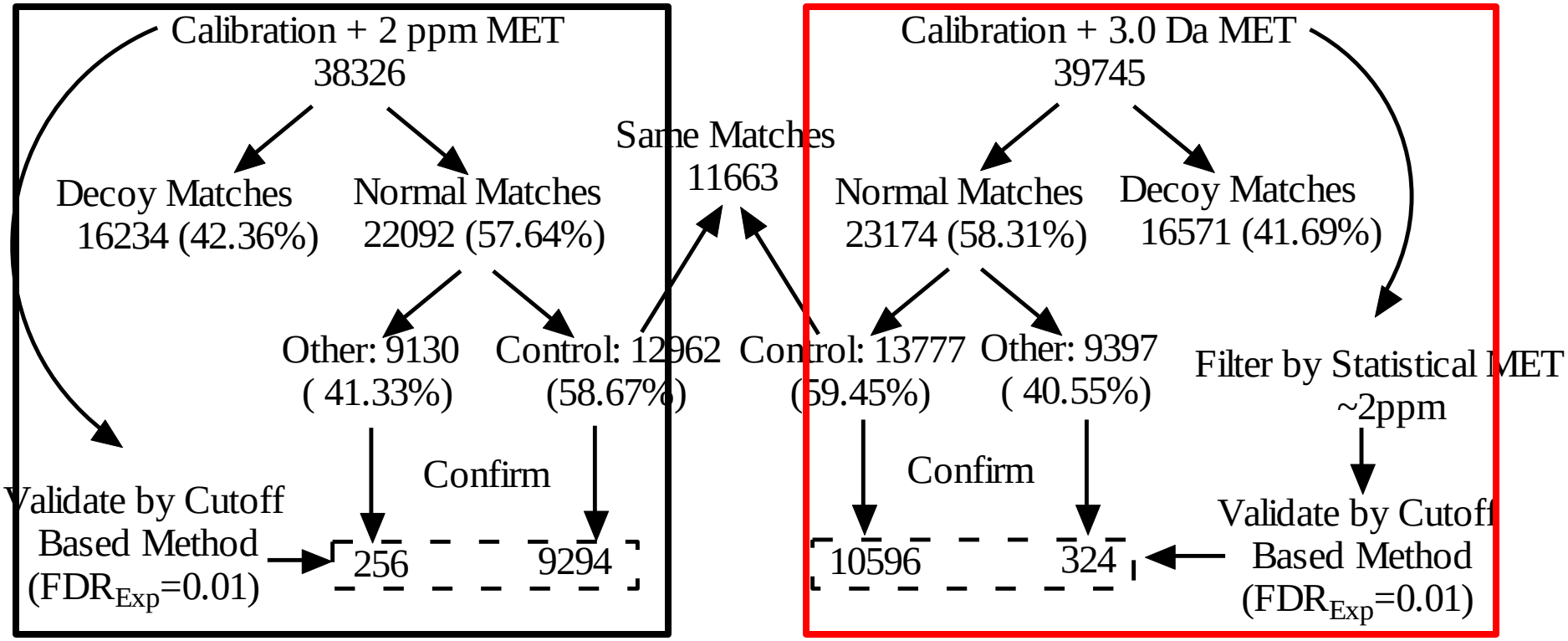
LDSF can improve the sensitivity of the result validation procedure



The preserved regions of the database search results of the control protein data set using mass calibration with small-MET (left) and large-MET (right) strategies.

- The database search scores become more powerful in distinguishing the peptide identifications and thus improve the sensitivity of the cutoff-based method.

LDSF can increase the validated peptide number



- Using LDSF strategy, we observed 10,920 validated peptides; this was 14.3% more than for the small MET database search, which yielded 9,550 validated assignments.

J Proteome Res. 2009 Feb;8(2):849-859.

Improvements of the filter methods of peptide identification in SEQUEST database search results

- A new strategy to filter out false positive identifications of peptides in SEQUEST database search results
- A nonparametric model for quality control of database search results in shotgun proteomics
- Bayesian nonparametric model for the validation of peptide identification in shotgun proteomics

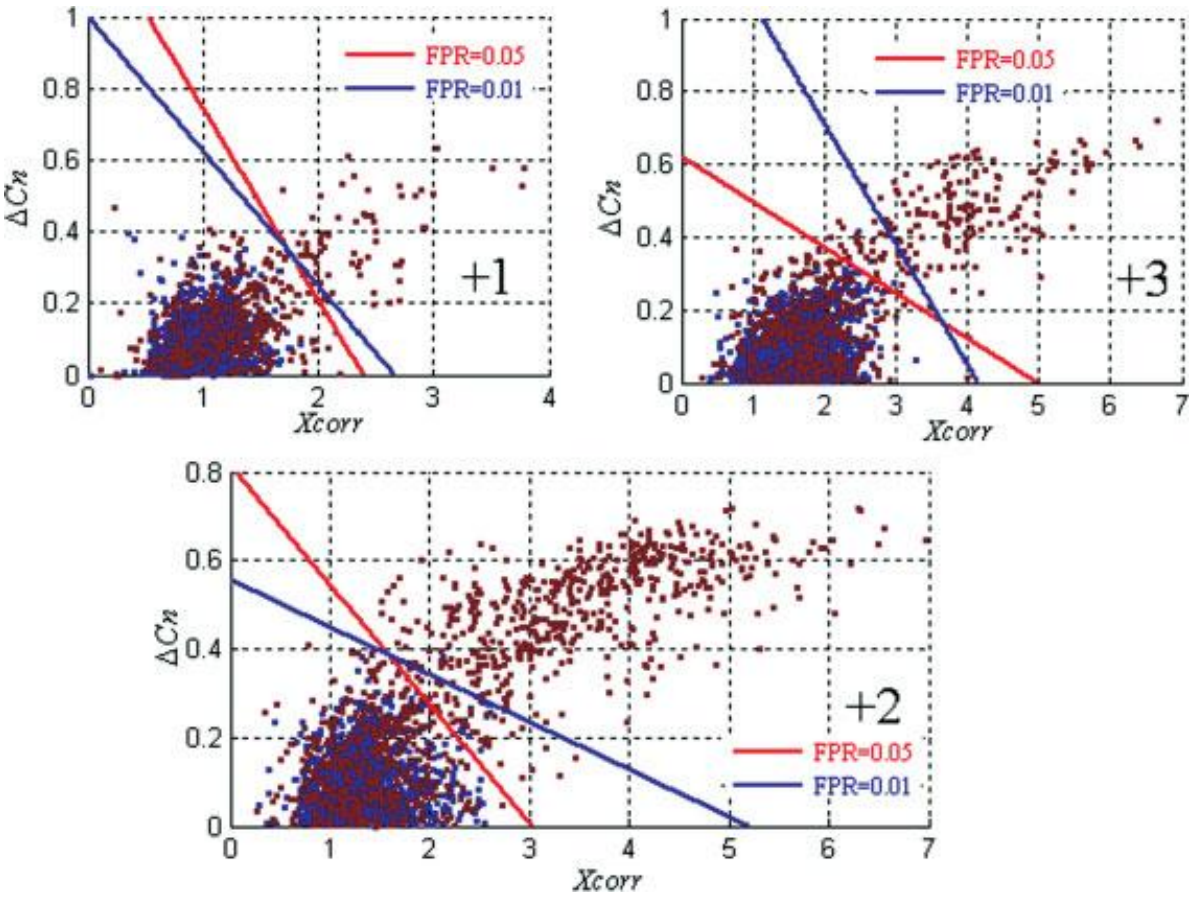
Peptide identified by SEQUEST

- Without any filtering, there would be many false positive assignments within the results of SEQUEST. (*J Am Soc Mass Spectrum.* 2002, 13(4):378-386. *Anal. Chem.* 2002, 74(21):5593–5599. *Mol Cell Proteomics*. 2004, 9(4):173~181.)
- Many works on the validation of SEQUEST database search results have been published, but each has its own shortage.
  - Empirical cut-off based method
    - Lack appropriate statistical foundations and good explanations
  - Probability models based methods
    - PeptideProphet
  - Machine learning models methods
    - Depend intensively on the quality of the selected characters as well as training set composition.
  - Randomized database based methods
    - It evaluate the quality of resulting dataset as a whole, which could not detect the accuracy of each assignment

A new strategy to filter out false positive identifications of peptides in SEQUEST database search results

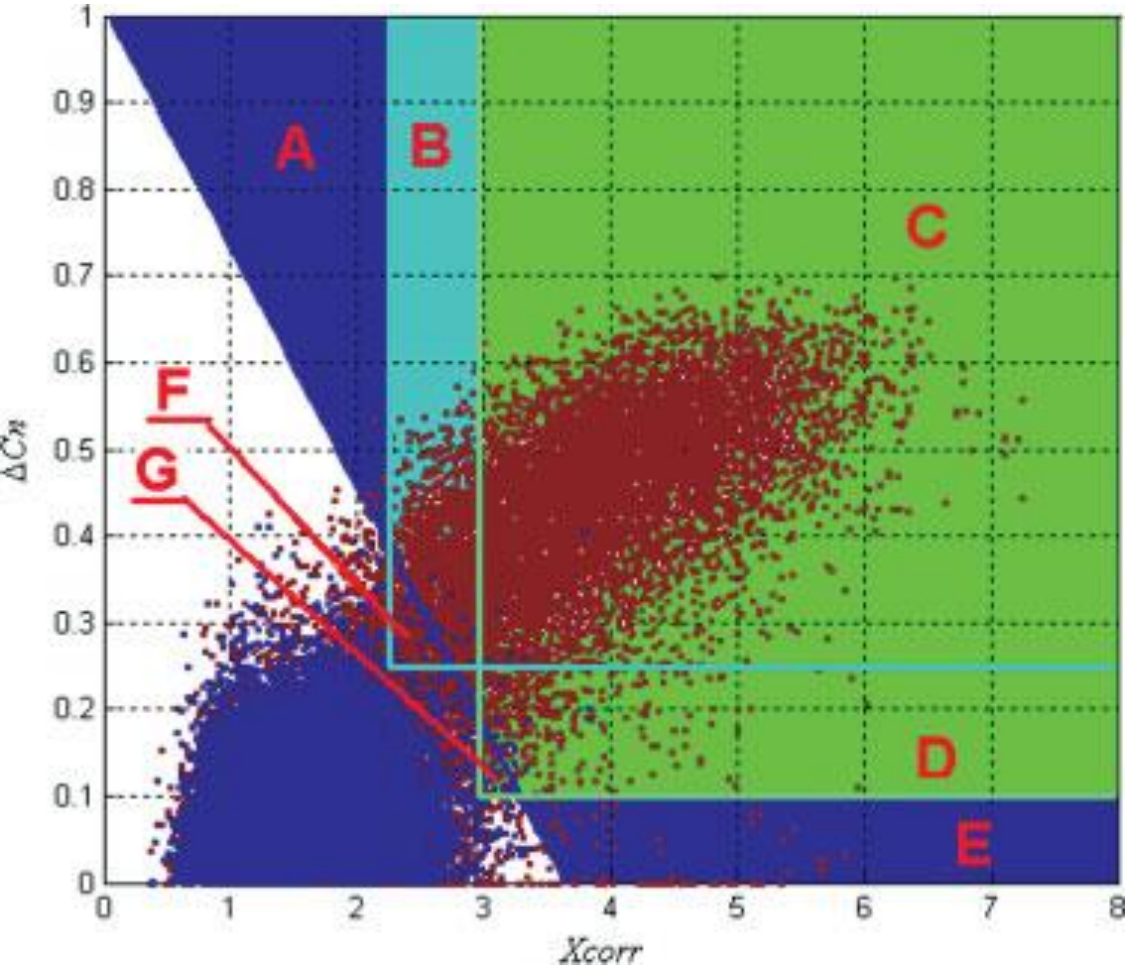
- Based on the randomized database method, a linear discriminant function (LDF) model is proposed to filter out false positive matches in SEQUEST database search results.
- The LDF model takes into account the dynamic tradeoff between *Xcorr* and Δ*Cn* through the use of a filtering boundary: Δ*Cn* = *k (b-Xcore)*.
- The coefficients (*k, b*) pairs are determined by keeping the *FDR* fixed and maximizing the number of normal database matches after filtration.

The filter boundaries derived from the LDF model



- The filtration was applied to the +1, +2, and +3 charge state data respectively
- The red and blue points are the normal and randomized database matches
- The red and blue line is the LDFs at FDR of 0.05 and 0.01.

Comparing preserving regions on *Xcorr*–Δ*Cn plane* determined by three filtering methods



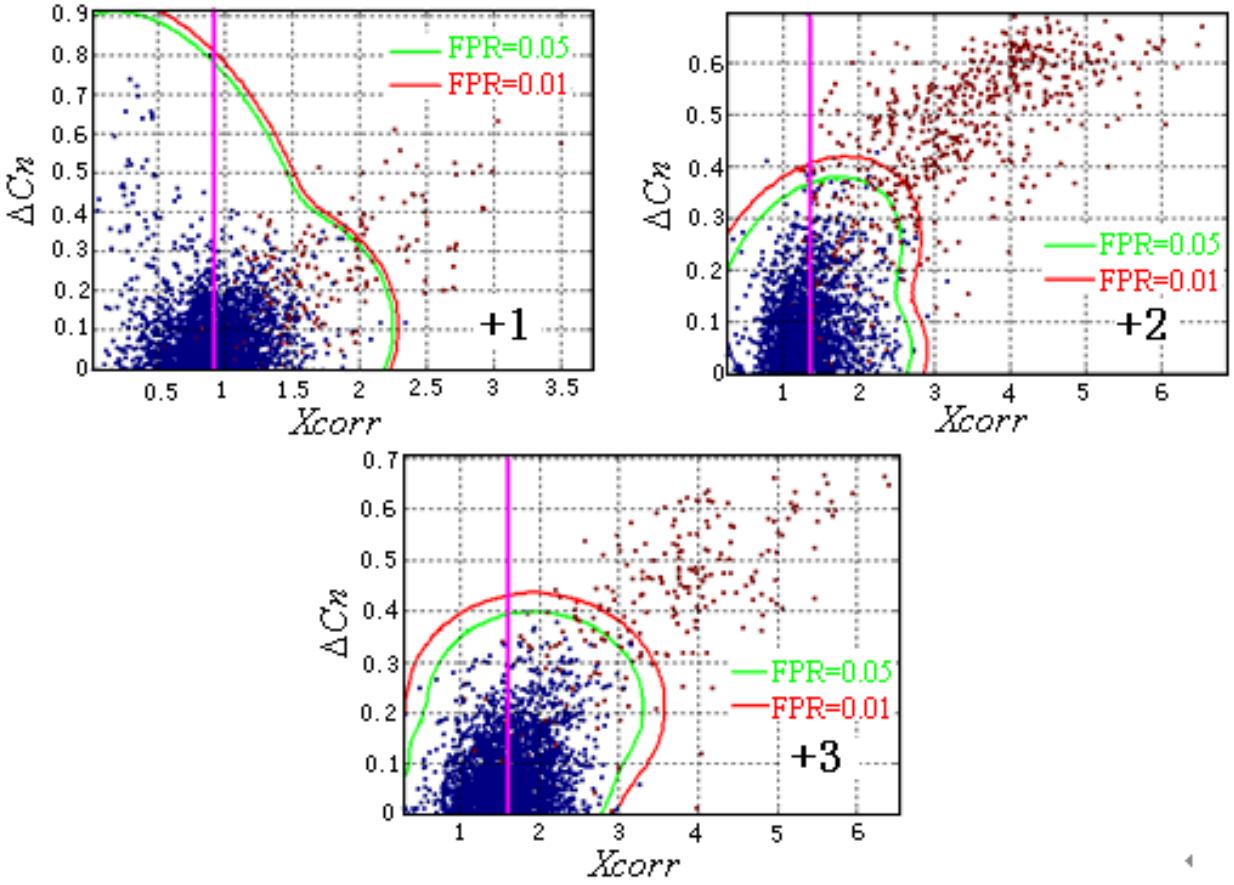
- The LDF model gives the largest acceptance regions
  - A, B, C, D, E, and a small unlabeled triangle in the center.
- Method 1: fixed Δ*Cn*
  - C, D, and G.
- Method 2: optimal *Xcorr* and Δ*Cn*
  - B, C, and F.

Proteomics. 2007 19;7(22):4036-4044.

A nonparametric model for quality control of database search results in shotgun proteomics

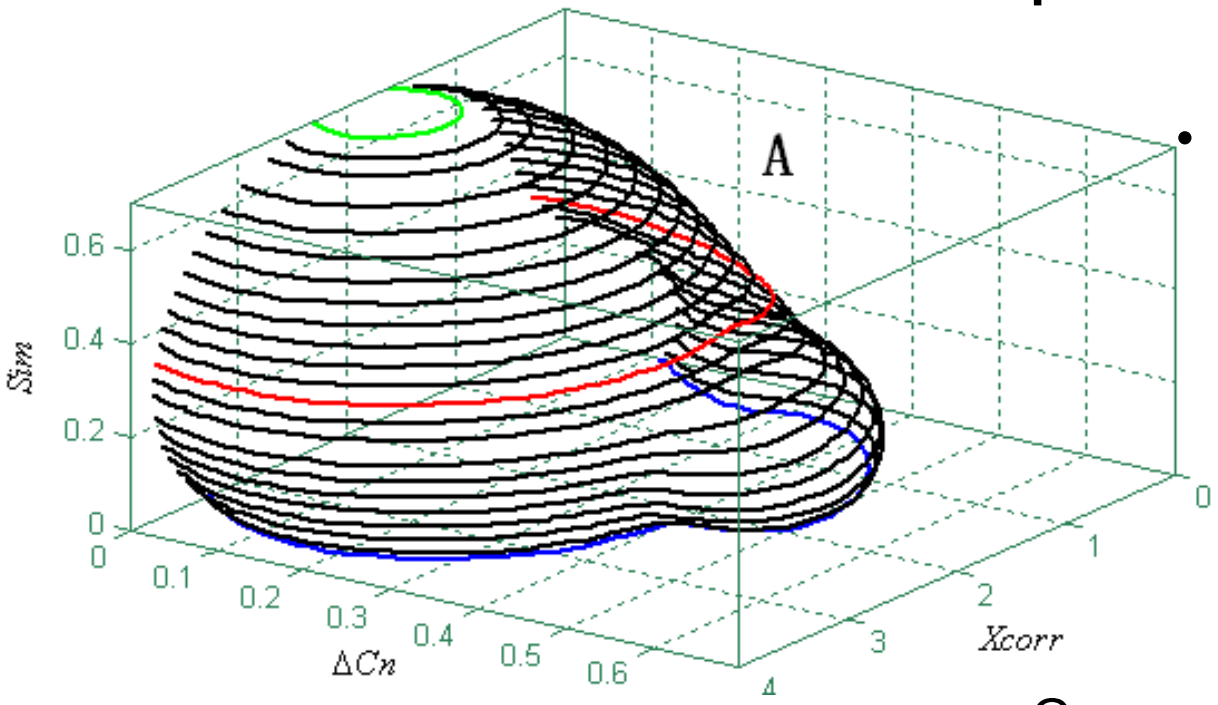
- The nonparametric model uses the nonparametric density estimation technique to estimate the distribution of the database search scores and takes the contour lines as the candidate discriminant functions to filter out false positive results.
  - More flexible: the number and nature of the parameters are not fixed in advance.
  - More accurate: the distribution of multiple parameters can be fit directly with considerable accuracy.
  - More sensitive: this nonparametric statistical technique is a powerful tool for tackling the complexity and diversity of datasets in shotgun proteomics.

Nonlinear filter boundaries on *Xcorr–ΔCn plane* by nonparametric model

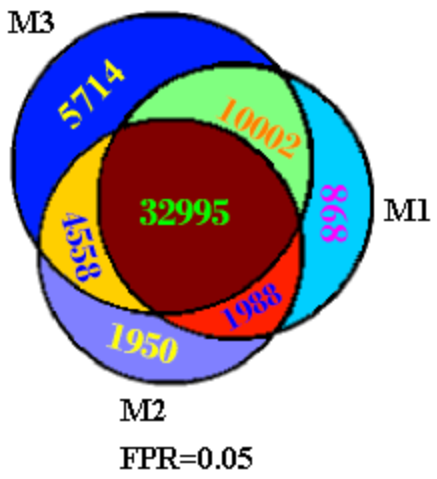


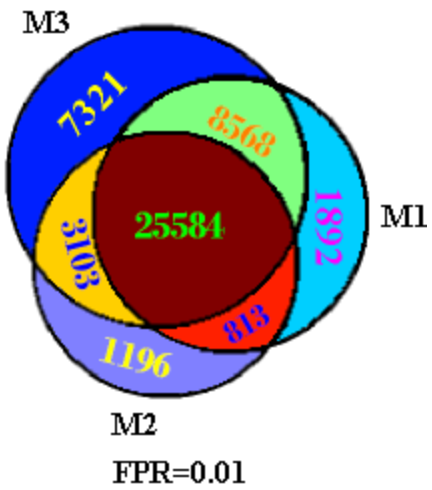
- Inferred filter boundaries for different charge state observations in the control dataset. The red and green curves are the filter boundaries for FPR = 0.01 and for FPR = 0.05, respectively.

Performing classification in a high-dimension feature space



- The nonparametric model can provide greater discriminating power by incorporating more features
  - *Xcorr*, △*Cn* and *SimScore*





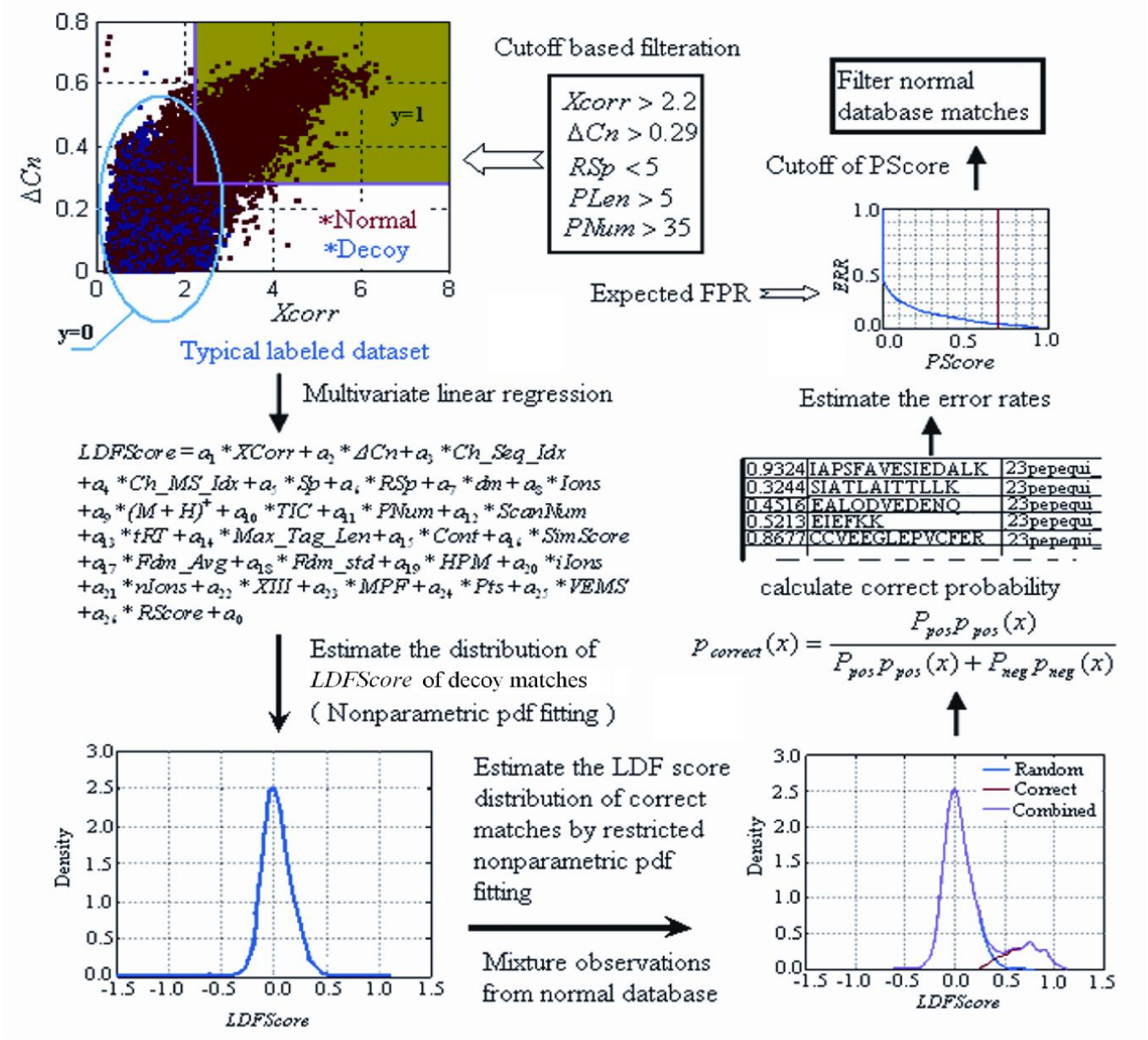
- Comparison of the confirmed matches among different method
  - M1, cutoff-based method
  - M2, peptideProphet
  - M3, nonparametric model
- The nonparametric model has the highest sensitivity

BMC Bioinformatics 2008, 9:29.

Bayesian nonparametric model for the validation of peptide identification in shotgun proteomics

- If too many parameters are used, the nonparametric model will encounter a computational problem.
- We developed a Bayesian nonparametric model (BNP) to filter the false positive matches in shotgun proteomics database searching.
  - Integrate a large number of features
  - Model the probability structure from the target–decoy database search results, and automatically classify the results
  - High power to separate correct from incorrect assignments
  - Greatly increase the number of confirmed peptides and proteins.

Workflow of Bayesian nonparametric model



The sensitivity of the BNP model surpassed that of three other filter methods

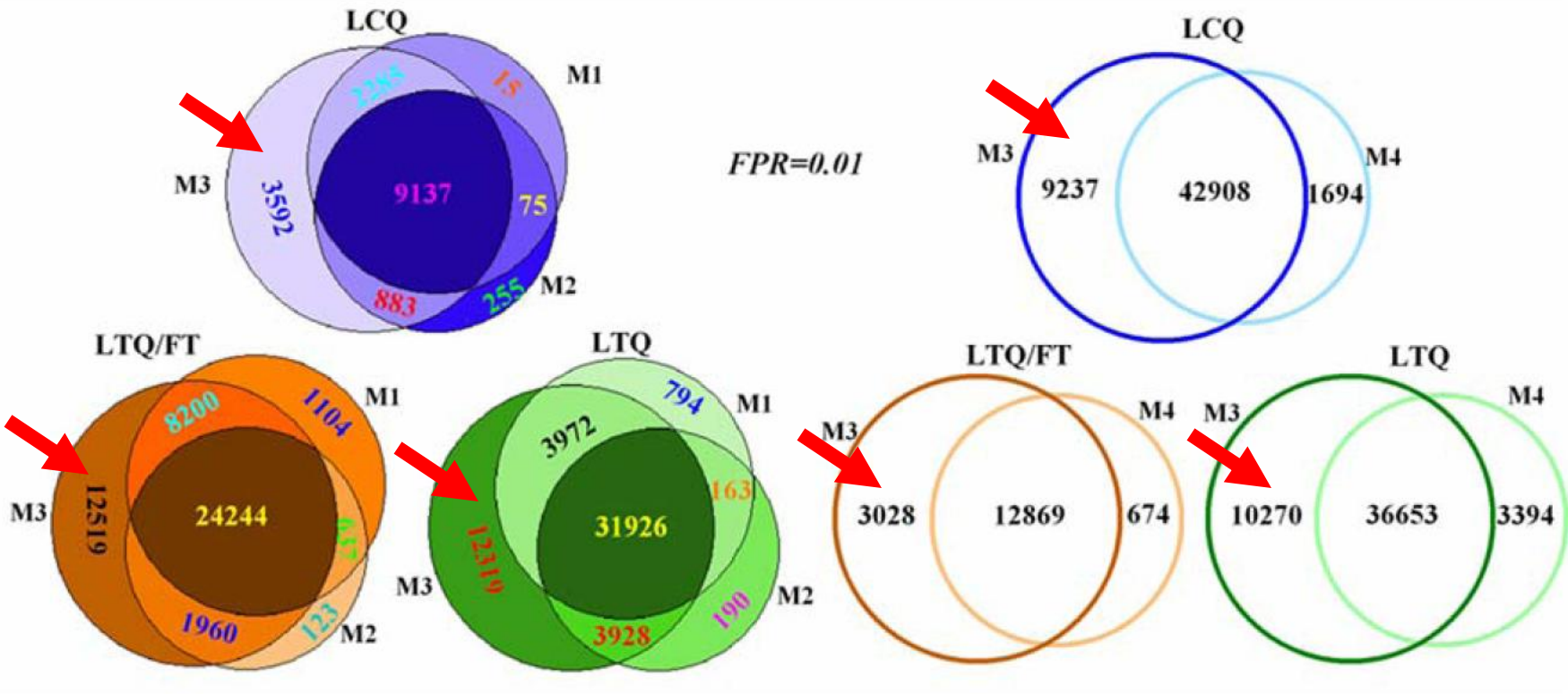
Comparison of four filter methods on control data sets

Dataset[a]	Method[b]	Expected FDR = 5%			Expected FDR = 1%		
		Actual FPR (%)	Total/Correct	Sensitivity (%)	Actual FPR (%)	Total/Correct	Sensitivity (%)
D1	M1	2.23	719/703	78.29	0.53	567/564	62.81
	M2	2.59	733/714	79.51	0.89	674/668	74.39
	M3	2.20	820/802	89.31	0.40	758/755	84.08
	M4	2.72	810/788	87.75	1.39	722/712	79.29
D2	M1	1.92	5875/5762	68.20	0.36	4964/4946	58.54
	M2	2.17	6775/6628	78.45	0.51	5895/5865	69.42
	M3	3.16	7426/7191	85.11	1.04	6754/6684	79.11
	M4	1.91	7001/6867	81.28	0.55	6333/6298	74.54
D3	M1	0.13	10284/10271	74.80	0.03	9182/9179	83.70
	M2	0.42	11477/11429	93.14	0.17	10699/10681	87.04
	M3	0.50	11983/11923	97.16	0.09	11388/11378	92.72
	M4	0.32	10885/10850	88.42	0.16	10117/10101	82.32

[a] D1: LCQ control dataset; D2: LTQ control dataset; D3: LTQ/FT control dataset
[b] M1: Cutoff-based method; M2: PeptideProphet; M3: BNP model; M4: Nonparametric model.

- Under the 1% expected FDR, the BNP model validated about 5% ~ 36% more peptides than other methods.

The peptides confirmed by the BNP model represented more than 90% of other three methods

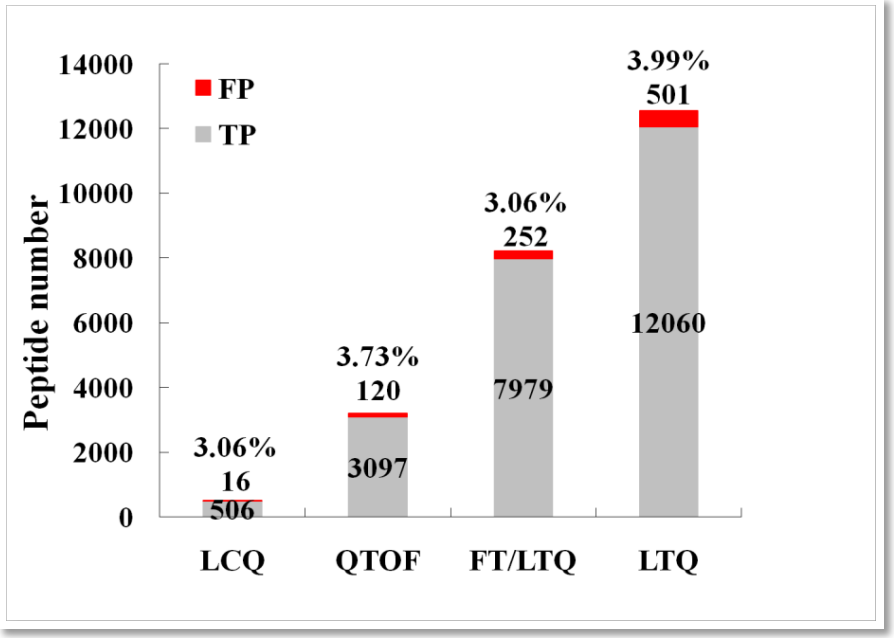


Overlap of peptides identified by the four methods.

M1: Cutoff based method; M2: PeptideProphet;
M3: BNP model; M4: Nonparametric model.

Peptide identified by Mascot

- Majority of the proposed filter methods for Mascot have been based on the ion score and thresholds reported by this search engine.
  - mass accuracy–based threshold (MATH) (*J. Proteome Res.* **2005,** 4, (4), 1353-1360.)
  - empirical Mascot homology threshold (MHT) (*Mol. Cell. Proteomics* **2008,** 7, (5), 962-970.)
  - transformed *E-value* (Biol. Direct **2007,** 2, (26).)
- The *.dat files output by Mascot contain extensive information.



- The Mascot identity threshold (MIT) can control the FDR of search results strictly.
  - low sensitivity
  - lost a great deal of true results

Combination of new features improves peptide identification by Mascot in shotgun proteomics

On the basis of target-decoy search strategy

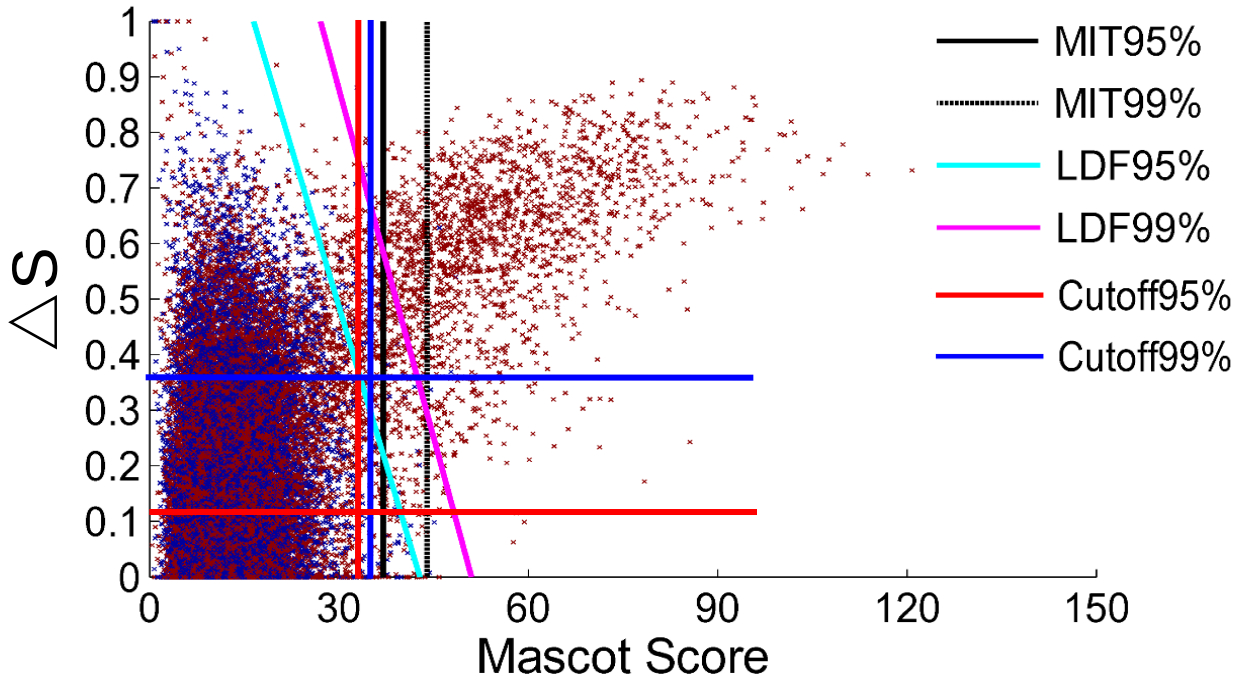
Introduce new features to improve the discriminant power of the Mascot search score

Apply robust filter methods to improve the sensitivity of result validation

Improve the filter process of Mascot search results

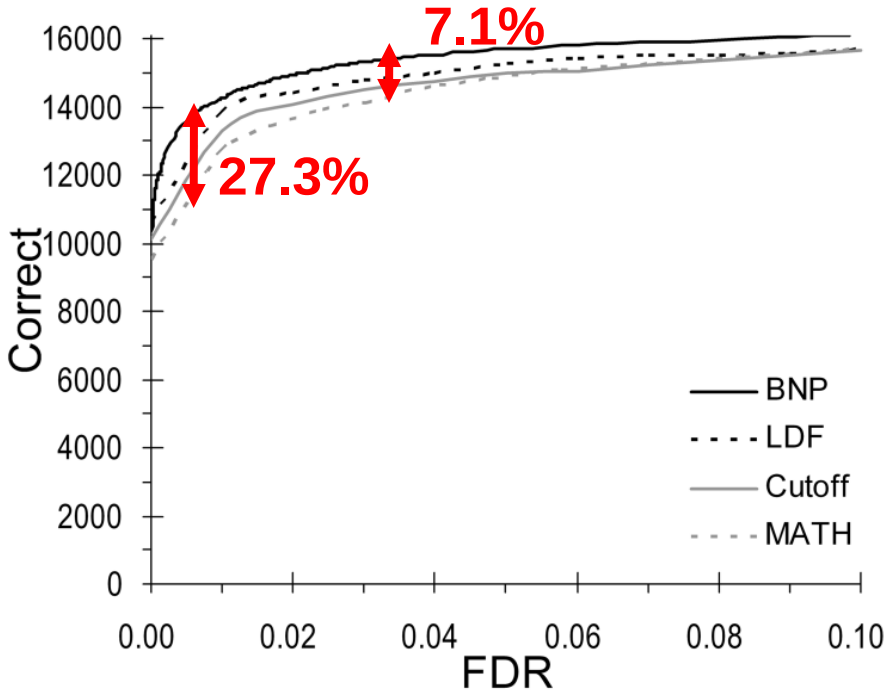
New features to improve the discriminant power of the Mascot search score

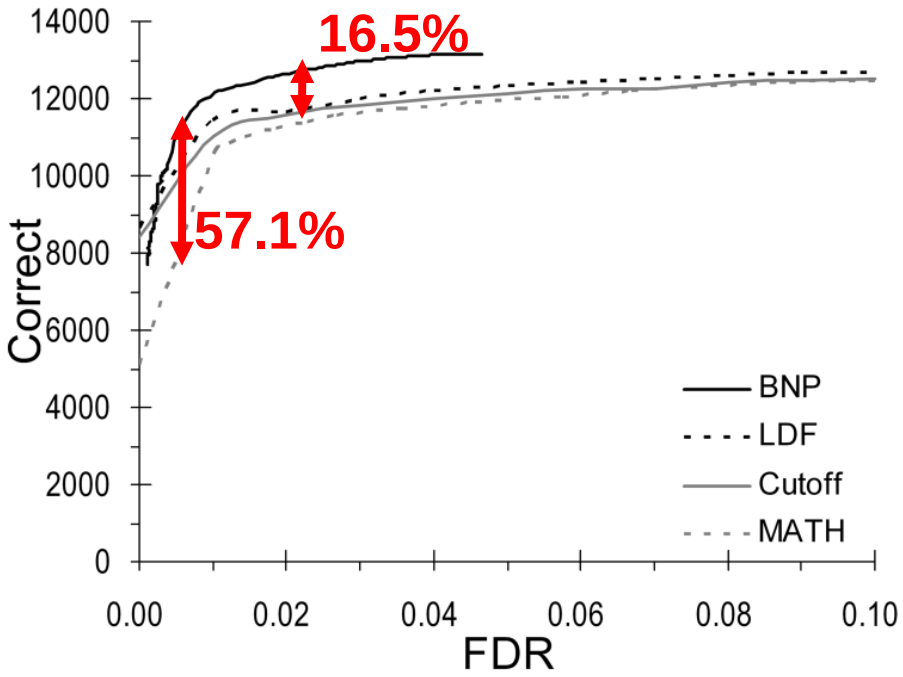
- Define the delta score for Mascot
  - $\Delta S = 1\text{-}Score_2/Score_1$
- Filtration of Mascot search results on the Score-ΔS plane
  - The cutoff based method
  - linear discriminate function (LDF) filtering boundary



Filter boundaries derived from Mascot identity threshold (MIT), cutoff-based method, and LDF model for LTQ-FT control data set

Application of BNP model to validate Mascot database search results





Plot figures of the number of correctly identified peptides *vs.* the estimated FDR on LTQ control data set (D2, left) and FT/LTQ control data set (D4, right).

- A total of 28 features are combined by BNP model to improve the validation of database search results for Mascot searched files
- The BNP model validated more correctly identified peptides than the other three methods.

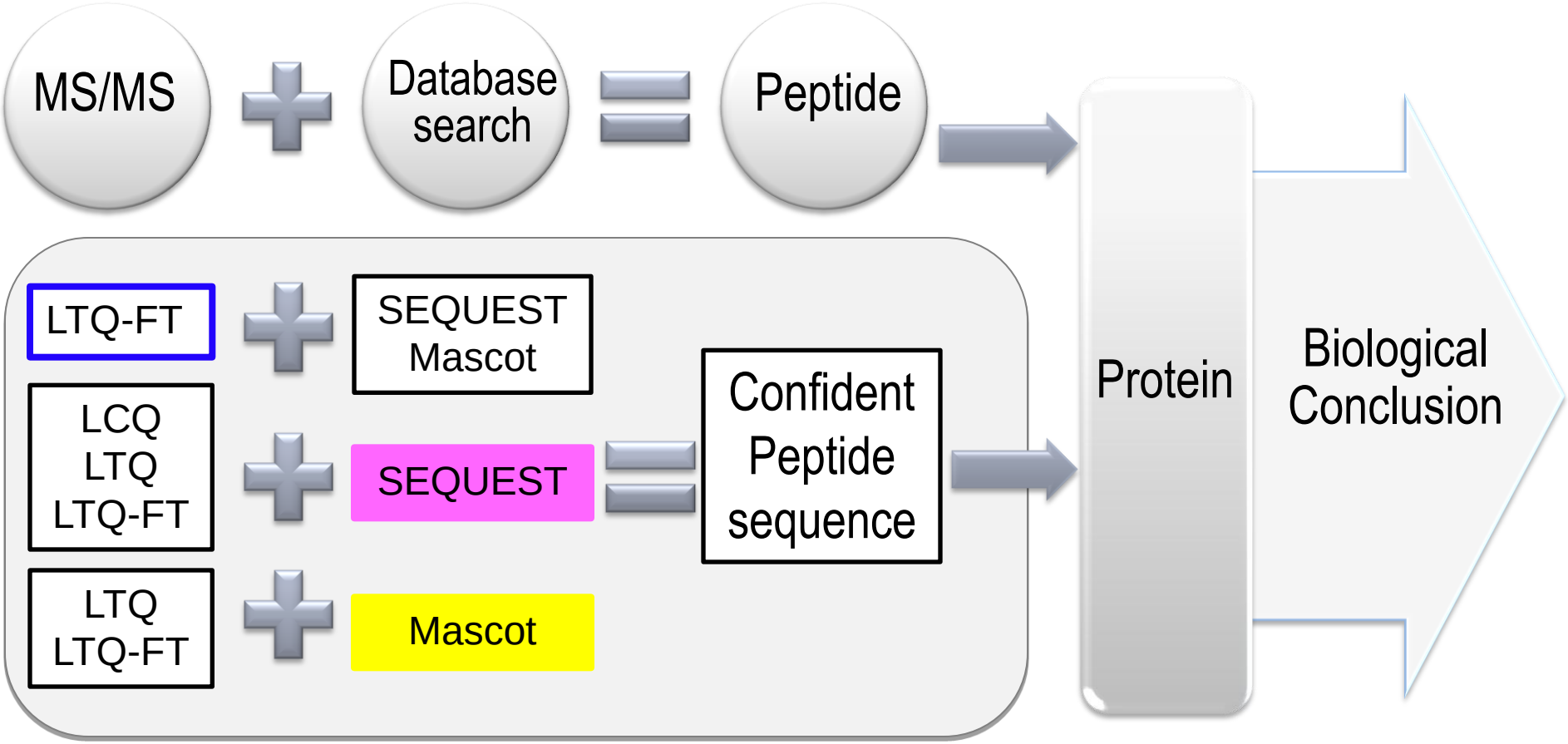
Comparative evaluation of filter methods on complex data sets

Data set	Method[a]	Expected FDR = 5%		Expected FDR = 1%	
		Confirmed Peptides / More than MIT (%)	Non-redundant Peptides / More than MIT (%)	Confirmed peptides / More than MIT (%)	Non-redundant peptides / More than MIT (%)
LTQ	M1	12,309/0.00	2,071/0.00	10,199/0.00	1,787/0.00
	M2	16,586/34.75	2,824/36.36	14,375/40.95	2,425/35.70
	M3	16,931/37.55	2,903/40.17	14,462/41.80	2,440/36.54
	M4	17,144/39.28	2,964/43.12	14,683/43.97	2,465/37.94
	M5	19,311/56.89	3,458/66.97	16,632/63.07	2,777/55.40
LTQ-FT	M1	75,041/0.00	4,636/0.00	58,261/0.00	3,741/0.00
	M2	102,270/36.29	5,829/25.73	79,359/36.21	4,559/21.87
	M3	101,337/35.04	10,211/120.25	73,999/27.01	5,994/60.22
	M4	101,413/35.14	10,230/120.66	78,275/34.35	6,699/79.07
	M5	117,886/57.10	11,812/154.79	88,459/51.83	6,966/86.21

[a]: M1, MIT; M2, MATH; M3, cutoff - based method; M4, LDF model; M5, BNP model.

- Determination of the filtering criteria for the *Score*-Δ*S* two dimension feature space was more sensitive than that for the Mascot score;
- The BNP model yielded approximately up to 64% more total results than Mascot threshold methods

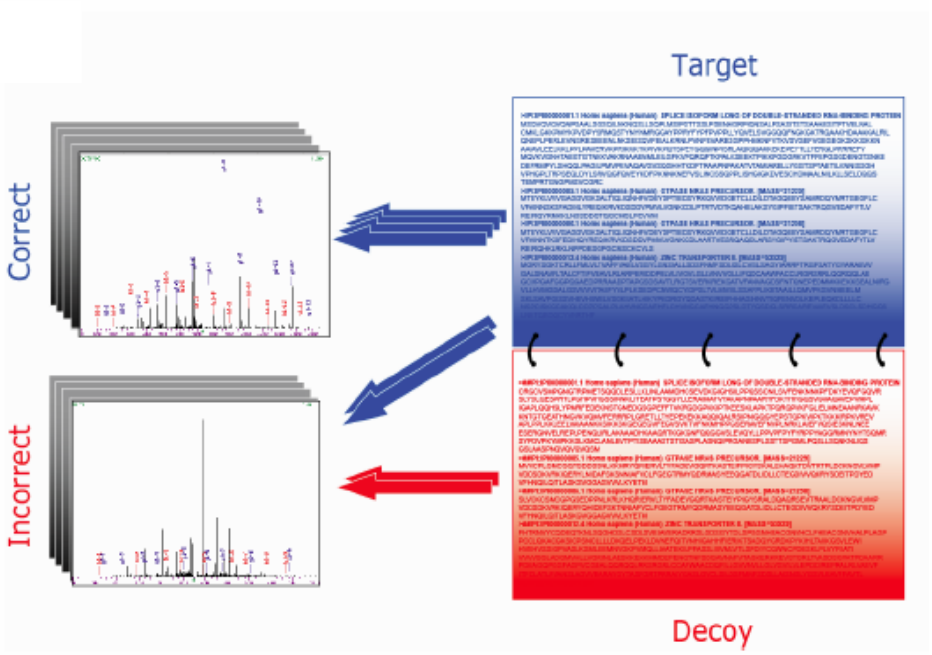
Summary

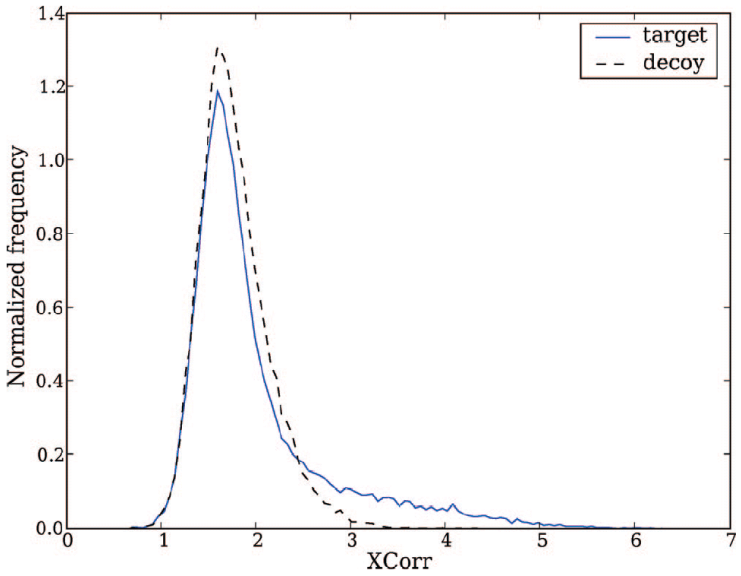


Quality control of Peptide identification in shotgun proteomics

Evaluation of the effects of decoy design, search strategy, and mass tolerance on the accuracy and sensitivity of peptide identifications in shotgun proteomics

Target-decoy search strategy





$$E\{FDR(s)\} = \frac{\pi_0 \frac{m_t}{m_d} |\{d_i > s, i = 1,2,...,m_d\}|}{|\{t_i > s, i = 1,2,...,m_t\}|}$$

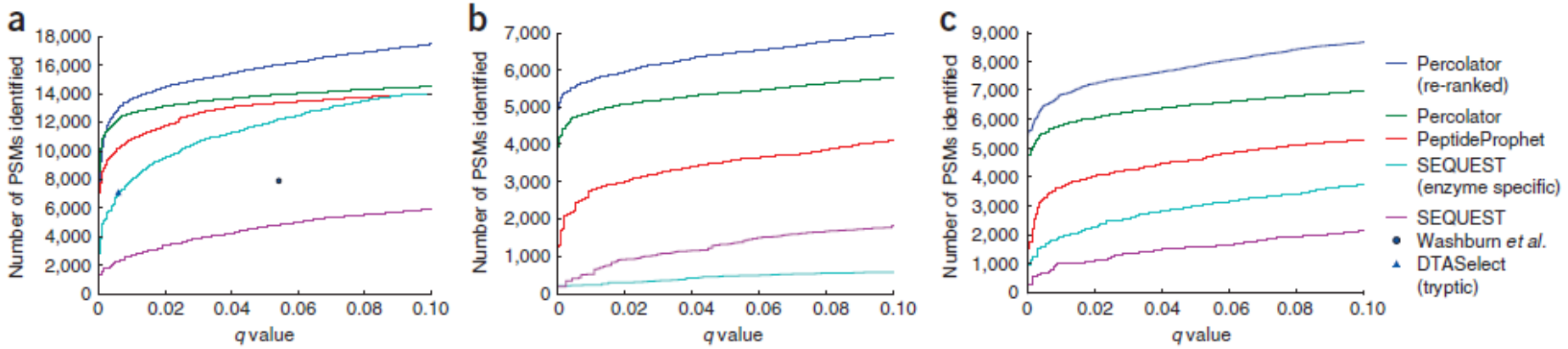
$$q(s) = \min_{s' \le s} E\{FDR(s')\}$$

Hypothesis： **Incorrect Peptide-Sequence-Matches(PSMs) from target or decoy sequences are equally likely.**

Elias, J.E. and S.P. Gygi, ***Nat Methods***, *2007. 4(3): p. 207-214*

Kall, L., et al., ***Nat Methods***, *2007. 4(11): p. 923-925*

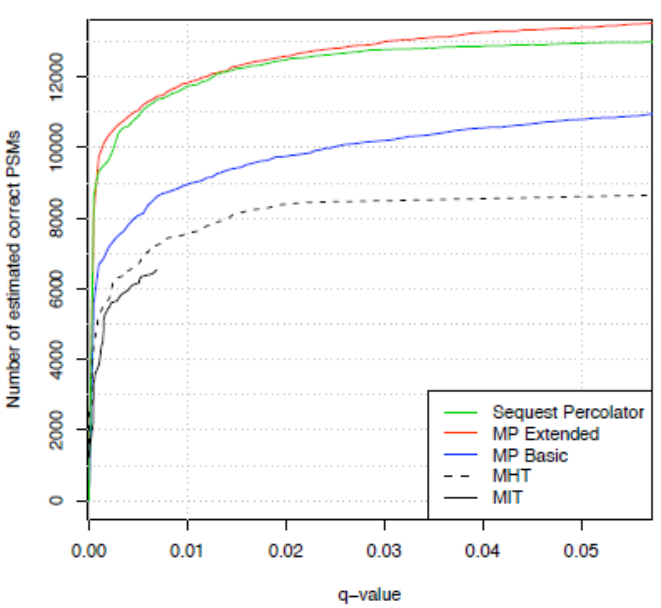
Percolator- Improve the sensitivity of peptide identifications



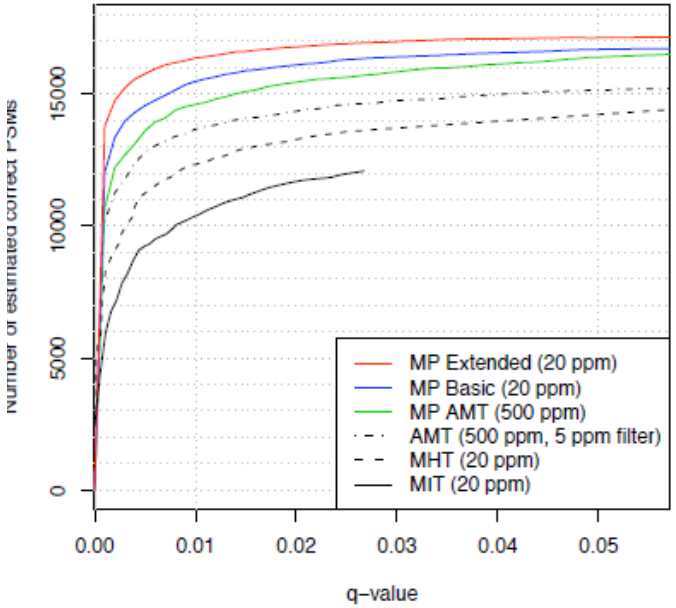
SEQUEST, LTQ, trypsin

SEQUEST, LTQ, elastase

SEQUEST, LTQ, chymotrypsin



MASCOT, LTQ, trypsin



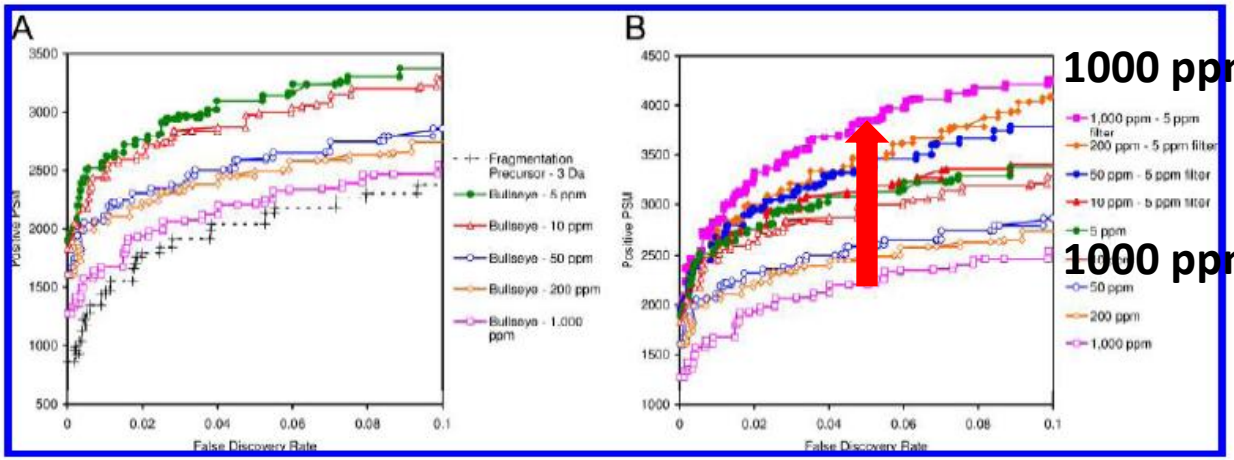
MASCOT, LTQ-FT, trypsin

q-value accuracy?

Kall, L., et al., ***Nat Methods**, 2007. 4(11): p. 923-925*

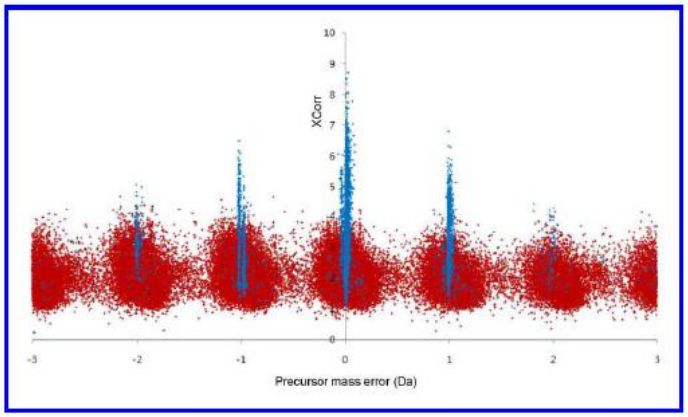
Brosch, M., et al., ***J Proteome Res**, 2009. 8(6): p. 3176-81.*

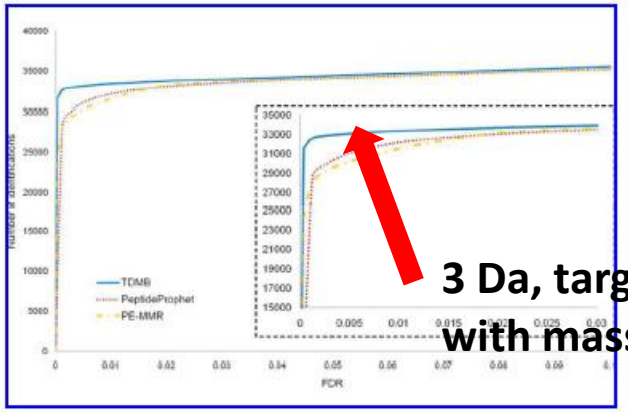
Wide precursor mass tolerance may improve the sensitivity of peptide identification



1000 ppm-5 ppm filter

1000 ppm





3 Da, target-decoy with mass binning

However, will the accuracy of FDR/*q*-value estimations be affected if we broaden the precursor mass tolerance?

Hsieh, E.J., et al., ***J Proteome Res**, 2010. 9(2): p. 1138-1143*

Joo, J.W., et al., ***J Proteome Res**, 2010. 9(2): 1150-1156*

Potential influencing factors

- Decoy design
- Search strategy
- Precursor mass tolerance
- Quality control (QC) method

We aim to find the appropriate decoy design, search strategy, and precursor mass tolerance to achieve both accurate and sensitive peptide identifications

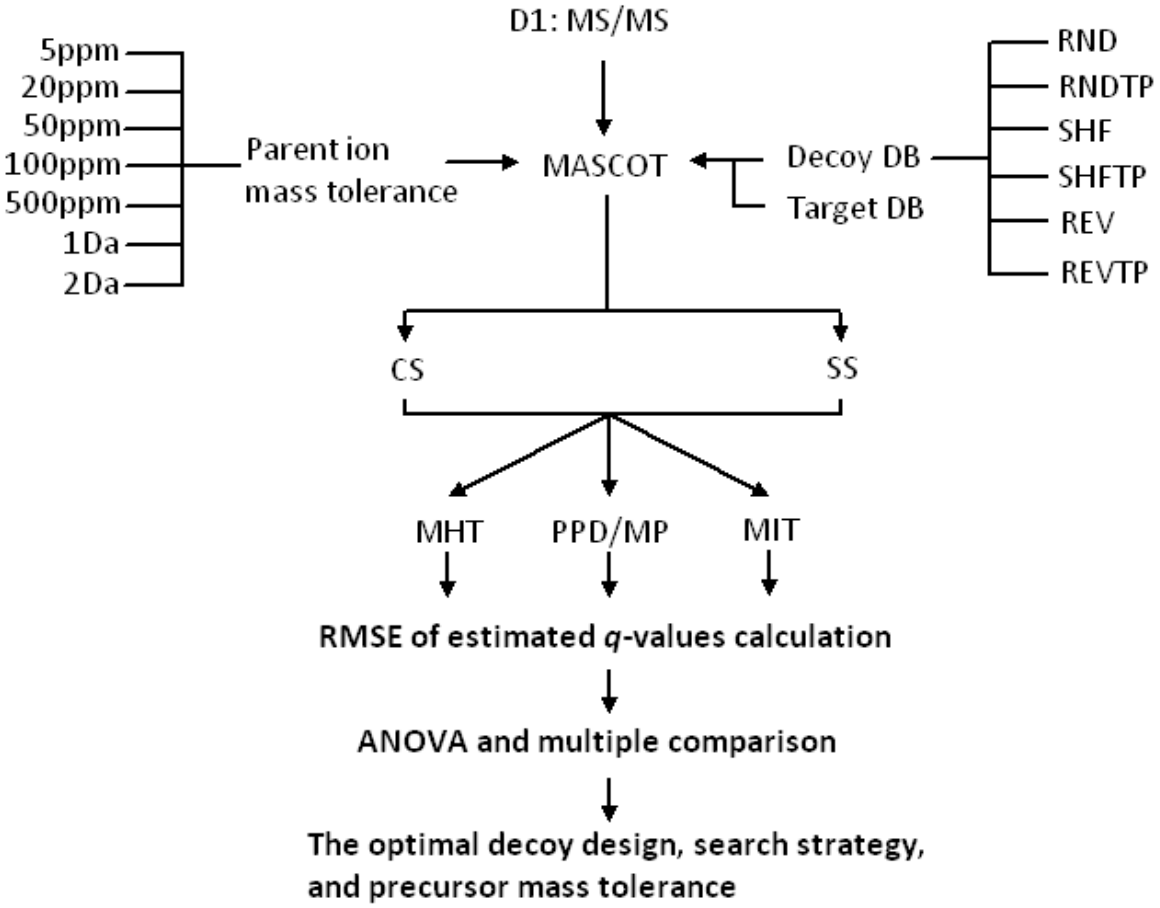
Blanco, L., J.A. Mead, and C. Bessant, **J Proteome Res**, 2009. 8(4): p. 1782-1791
Wang, G., et al., **Anal Chem**, 2009. 81(1): p. 146-159
Timm, W., et al., ***Anal Chem**, 2010. 82(10): p. 3977-3980*
Higdon, R., et al., ***OMICS**, 2005. 9(4): p. 364-379*

Datasets

- D1 (8191 spectra): a protein standard dataset comes from a set of 48 human proteins (Sigma, Universal Proteomics Standard Set UPS1). The sample was tryptic-digested. The raw data was generated by LTQ-FT mass spectrometry.
- D2 (24403 spectra): a complex sample dataset comes from human liver tissue. The sample was tryptic-digested, and analyzed by LTQ-FT.

Brosch, M., et al., ***J Proteome Res**, 2009. 8(6): p. 3176-3181*

The workflow of *q*-value accuracy evaluation



D1：203 MASCOT searches

Methods

- Influence factors
  - Decoy design

	Random	Shuffle	Reverse
Protein	RND	SHF	REV
Tryptic Peptide	RNDTP	SHFTP	REVTP

  - Search strategy
    - Separate *vs.* Composite search (**SS** *vs.* **CS**)
  - Precursor mass tolerance
    - 5ppm, 20ppm, 50ppm, 100ppm, 500ppm, 1Da, 2Da
  - QC method
    - Multiple features
      - **PPDistiller (PPD)：PTM and Peptide Distiller**
      - **MP**: MASCOT Percolator
    - Single feature
      - **MIT**: MASCOT identity threshold
      - **MHT**: MASCOT homology threshold
- Metrics evaluating the accuracy of *q*-value estimation

$$RMSE = \sqrt{\frac{1}{m_t} \sum_{i \in \{1,2,\ldots,m_t\}} (q_{est}(t_i) - q_{act}(t_i))^2}$$

- ANOVA and multiple comparison

PPDistiller (PPD)

- MP is limited to processing SS results
- PPDistiller
  - 36 features
  - Percolator
  - Process SS and CS results
- PPDistiller can generate more accurate *q*-values

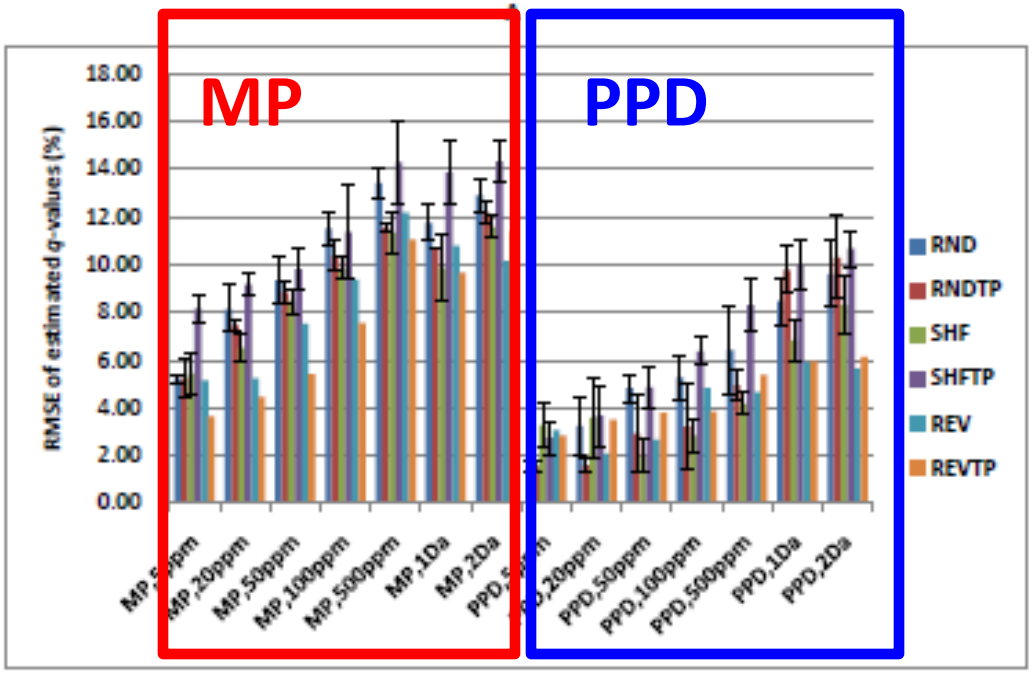


Figure S1A. The RMSEs of estimated *q*-values generated by MP and PPD for separate search results

PPD is better than MP for it can generate more accurate *q*-values, RMSE increases with the increase of precursor mass tolerances

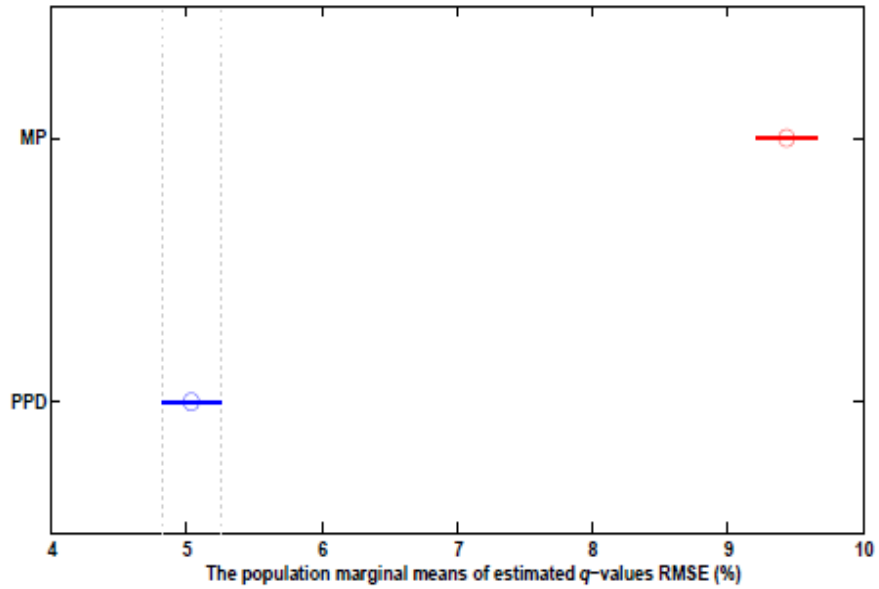


Figure S1B. Multiple comparison RMSE means of *q*-values generated by MP and PPD

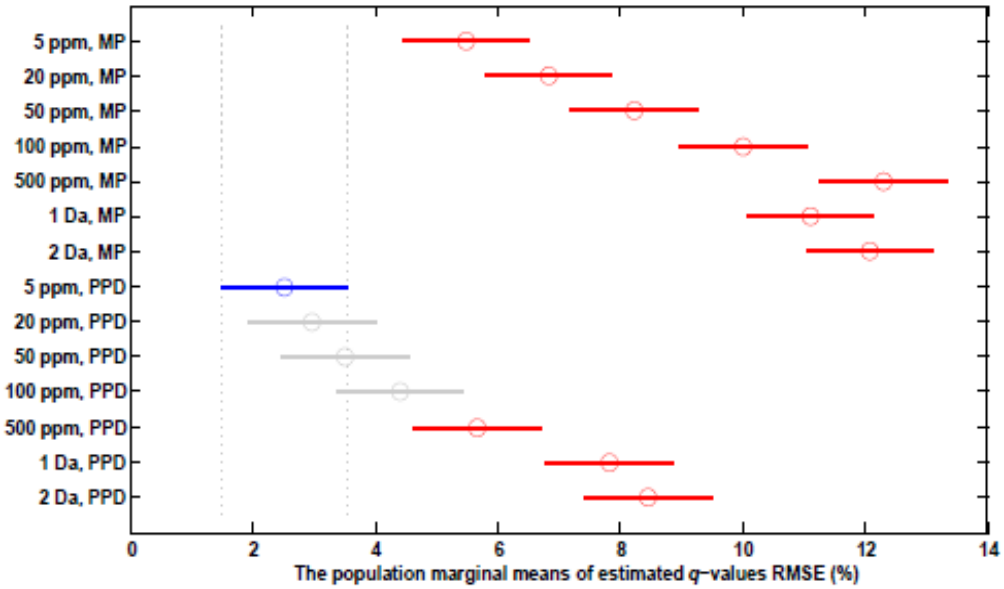
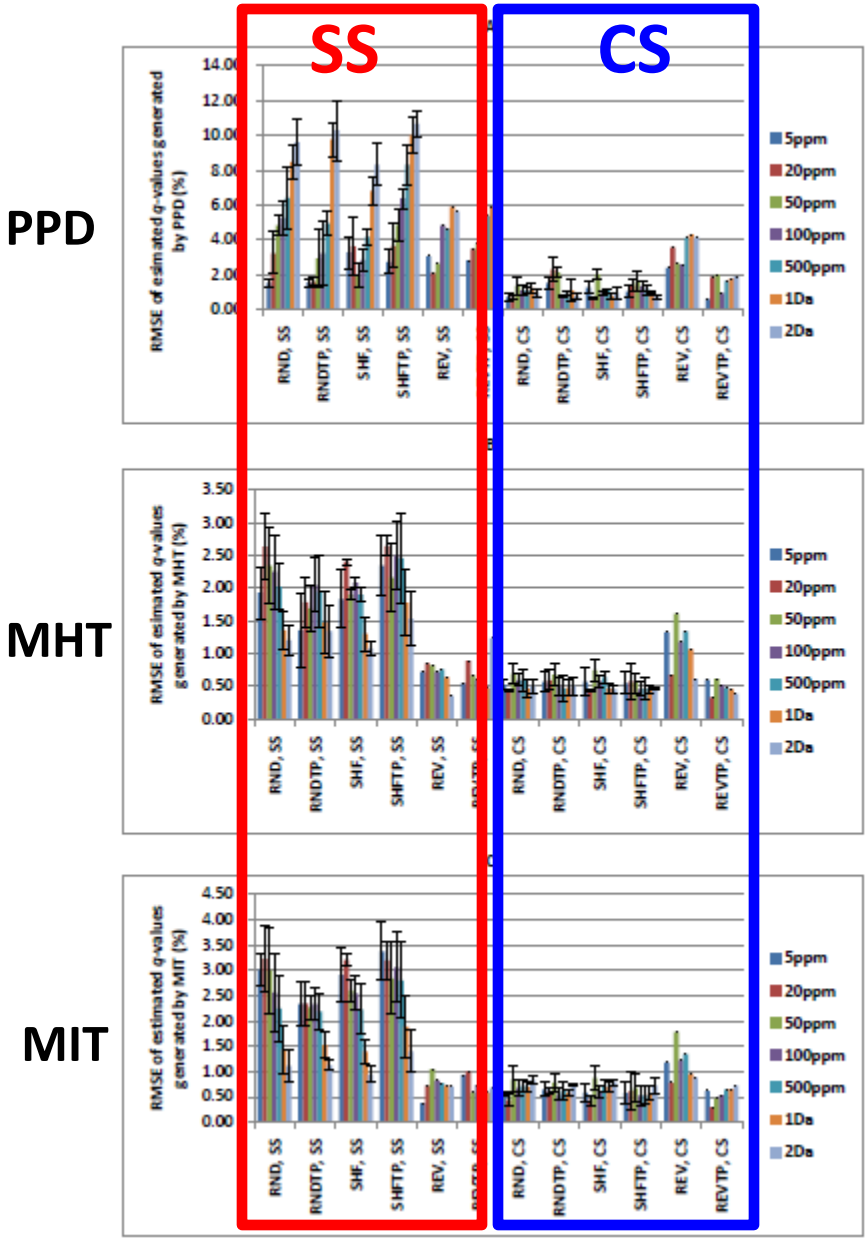


Figure S1C. Multiple comparison RMSE means of *q*-values for the 2-factor interaction: QC method*Precursor mass tolerance

The *q*-value RMSEs for PPD, MHT and MIT



- Decoy design
  - **For SS results, REV and REVTP were better than other decoy designs**
  - **For CS results, there was no significant difference between different decoy designs except for REV**
- Search strategy
  - **Composite search was better than separate search**
- **The decoy design and search strategy effects were reproducible across three QC methods**

Figure 1. RMSE means histograms of estimated *q*-values for various decoy designs, precursor mass tolerance settings, and quality control methods. The error bars represent standard deviations generated by repeatedly database searching for each copies of the stochastic decoy designs: RND, RNDTP, SHF, and SHFTP. A: the histogram of RMSE means generated by PPD; B: the histogram of RMSE means generated by MHT; C: the histogram of RMSE means generated by MIT.

ANOVA analysis

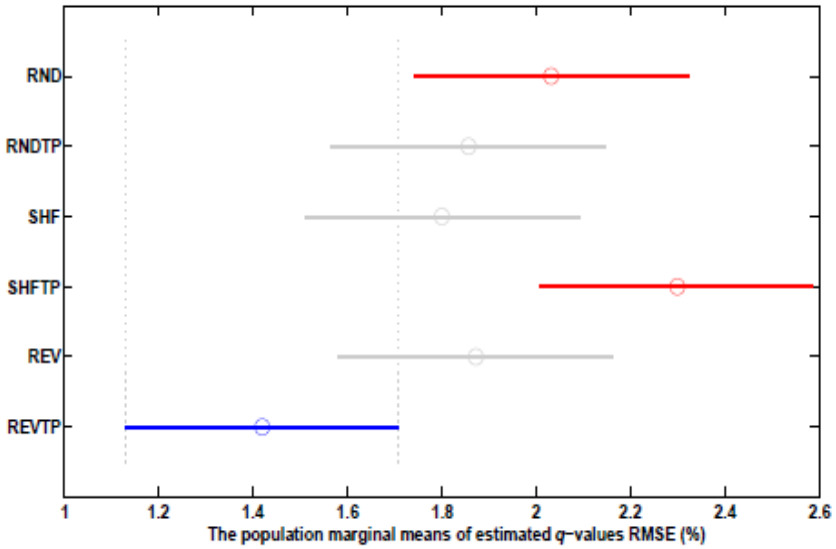
Analysis of Variance of RMSE of esimated q-values generated by PPD,MHT,MIT(%), 2-factor interaction effects

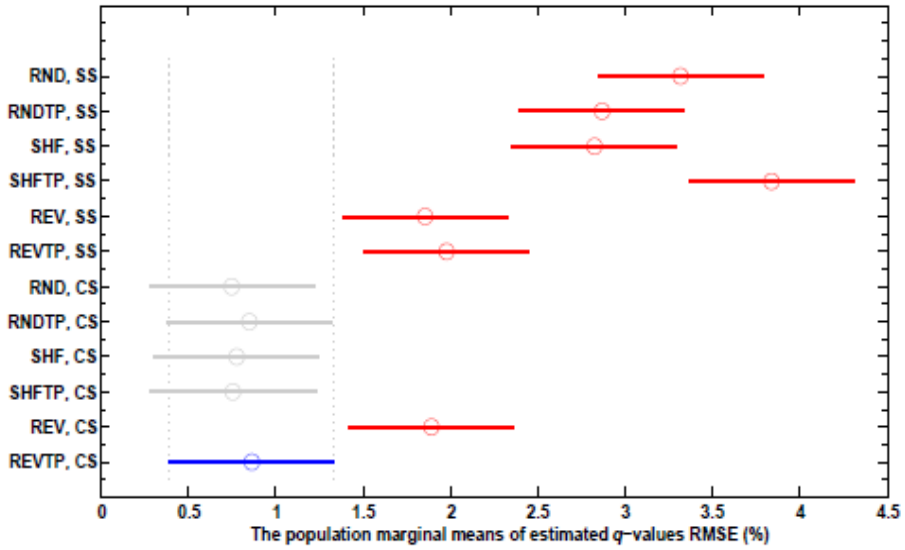
Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
DecoyDesign	17.55	5	3.51	4.02	**0.00**
ParentMassTolerance	17.74	6	2.96	3.39	**0.00**
SearchStrategy	203.83	1	203.83	233.74	**0.00**
QualityControl	262.35	2	131.18	150.42	**0.00**
DecoyDesign*ParentMassTolerance	5.26	30	0.18	0.20	1.00
DecoyDesign*SearchStrategy	64.93	5	12.99	14.89	**0.00**
DecoyDesign*QualityControl	11.54	10	1.15	1.32	0.22
ParentMassTolerance*SearchStrategy	17.91	6	2.99	3.42	**0.00**
ParentMassTolerance*QualityControl	91.41	12	7.62	8.74	**0.00**
SearchStrategy*QualityControl	85.42	2	42.71	48.98	**0.00**
Error	149.99	172	0.87		
Total	927.93	251			

Decoy design, search strategy, and precursor mass tolerance significantly affected the *q*-value accuracy

Most of the two-factor interactions significantly affected the *q*-value accuracy

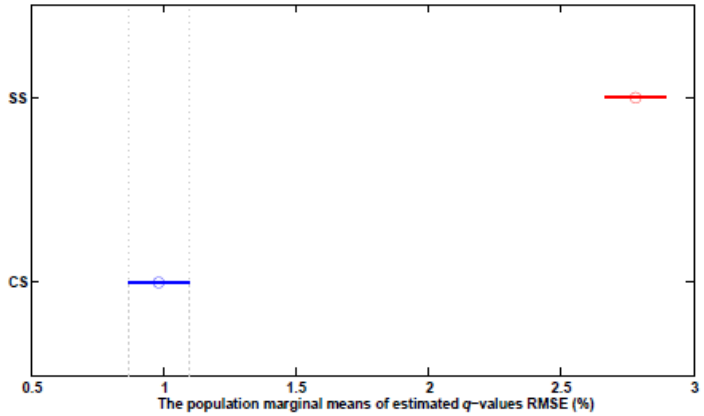
Accuracy: decoy design

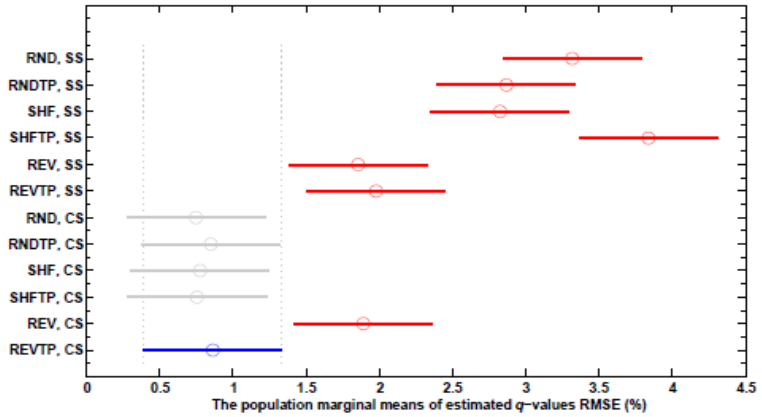


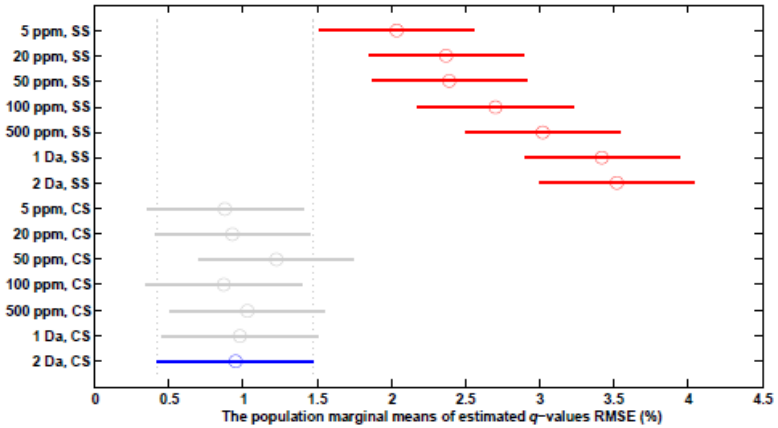


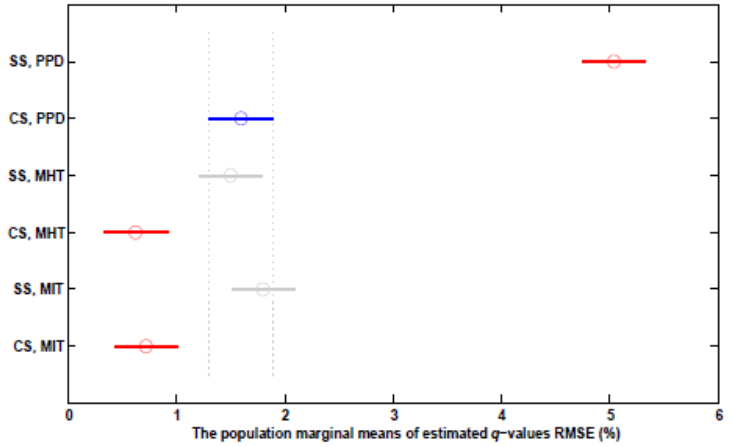
- REVTP was significantly better than RND and SHFTP
- Decoy design*Search strategy
  - For SS results, REV and REVTP was significantly better than the stochastic methods (i.e. RND, RNDTP, SHF and SHFTP)
  - For CS results, except for REV, there was no significant difference between the other five decoy designs

Accuracy: search strategy



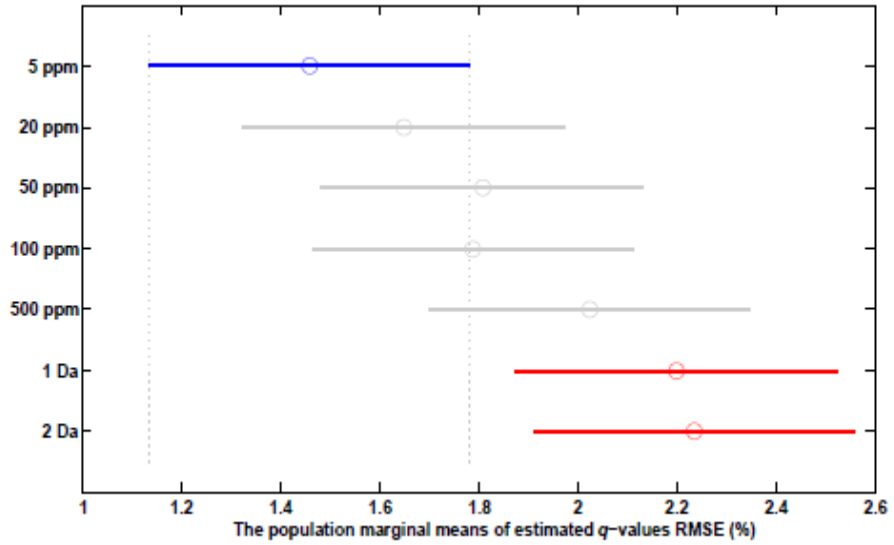


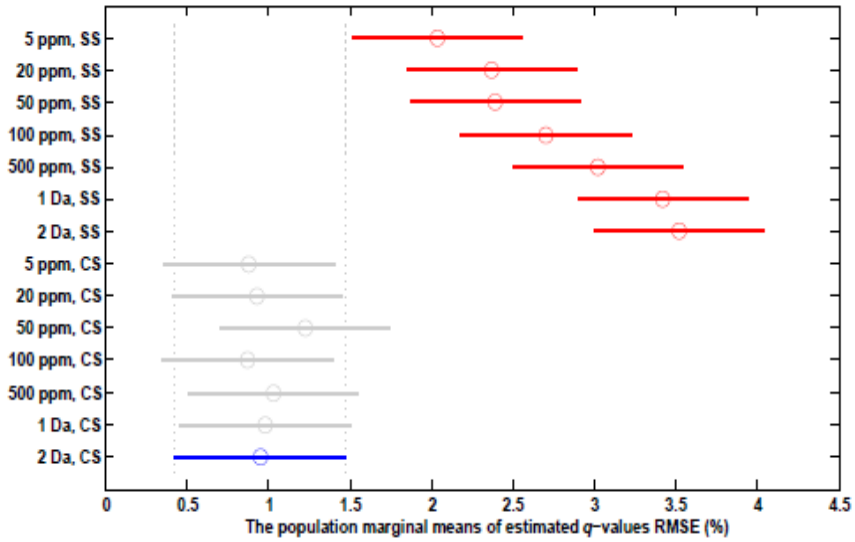


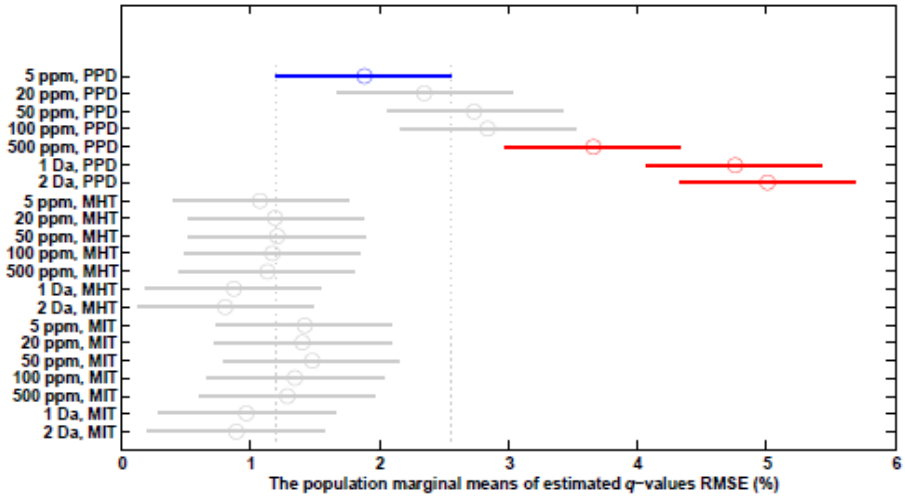


- CS was significantly better than SS . This effect was reproducible across different decoy designs (except for REV), mass tolerances and QC methods
- CS minimized the differences between different decoy designs, and eliminated the differences between different mass tolerances

Accuracy: precursor mass tolerance







- Narrow mass tolerance generated more accurate *q*-value estimations, specially for SS and PPD
- For CS, precursor mass tolerance didn't affected the *q*-value accuracy
- For MHT and MIT, precursor mass tolerance didn't affected the *q*-value accuracy

Summary
q-value accuracy evaluation

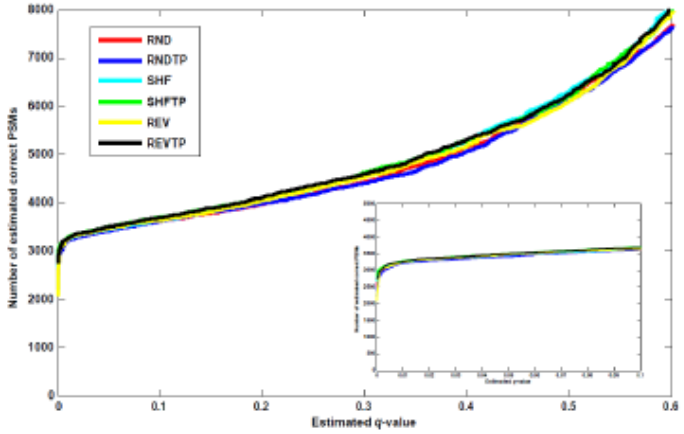
- **Composite search is better than separate search**. It can minimize or eliminate the differences between different decoy designs and precursor mass tolerances
- **For composite search, except for REV, the other five make no difference.** In separate search, REV and REVTP are better than the stochastic methods
- **For composite search, precursor mass tolerance doesn't affected the *q*-value accuracy;** For SS and PPD, narrow precursor mass tolerance can generate more accurate *q*-values

The sensitivity comparison

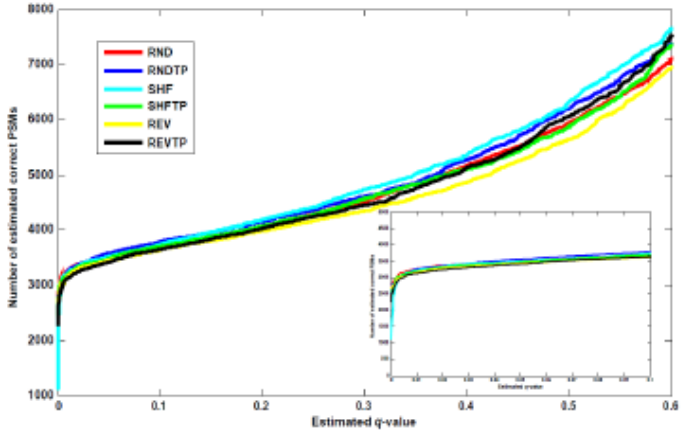
- D2 (24403 spectra)
  - Human liver tissue
  - Tryptic-digested
  - LTQ-FT
- QC method: PPD, MP
  - **Decoy design**
  - **Search strategy**
  - **Precursor mass tolerance**

Sensitivity: decoy design

A **MP**



B **PPD, SS**



C **PPD, CS**

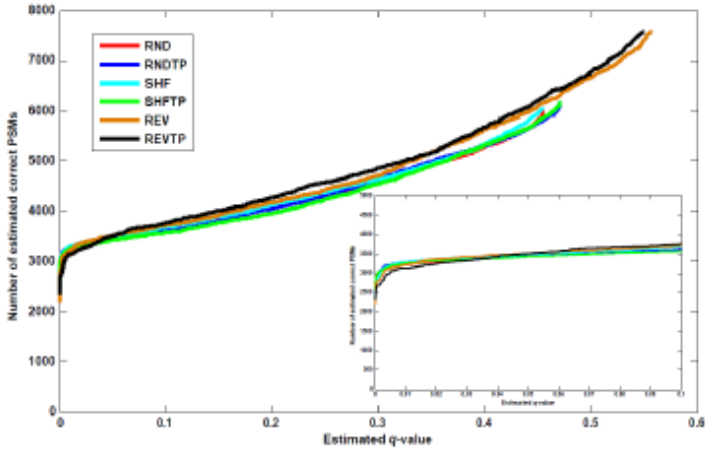


Figure 4. Sensitivity comparison of peptide identifications generated from different decoy designs. The dataset is HLFT. The number of estimated correct PSMs was plotted against each *q*-value threshold generated by (A) MASCOT Percolator (MP), (B) PPDistiller for separate search and (C) PPDistiller for composite search.

For MP and PPD (either in SS or CS mode), **different decoy designs achieved similar or at least compatible sensitivity, especially when *q*-value < 0.1**

Parent mass tolerance： 20 ppm

Sensitivity: search strategy

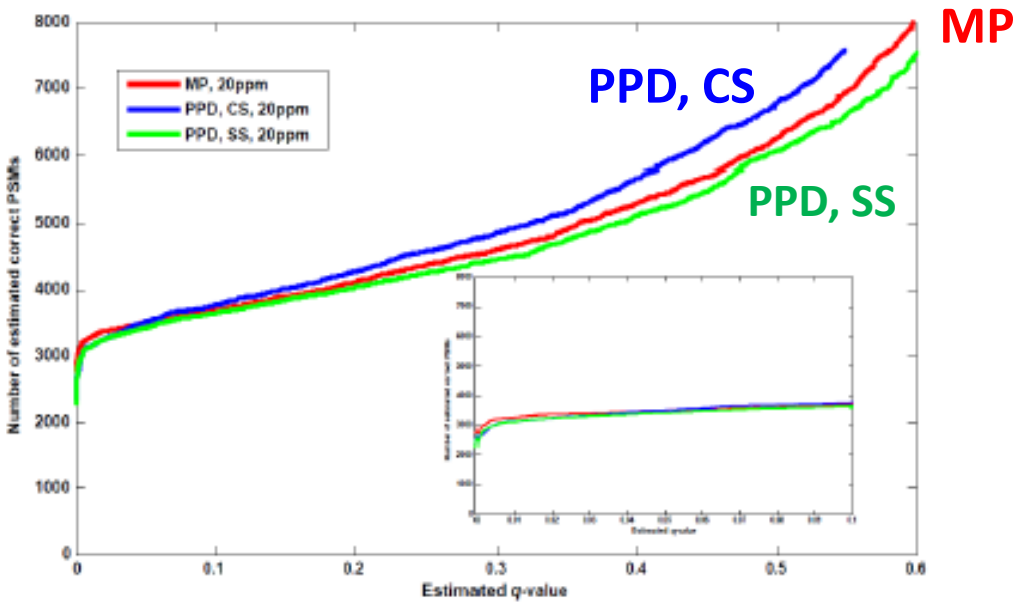


Figure 5. Sensitivity comparison of peptide identifications generated from composite and separate search. The parent mass tolerance was set as 20 ppm. Descoy database was designed using REVTP method. *q*-values were derived from MASCOT Percolator (MP), PPDistiller in composite search and separate search.

- PPD achieved similar sensitivity with MP
- More peptide identifications were obtained from CS results when *q*-value > 0.1;
- When *q*-value < 0.1, CS and SS made no difference

Decoy design： REVTP
Parent mass tolerance： 20 ppm

Sensitivity: precursor mass tolerance

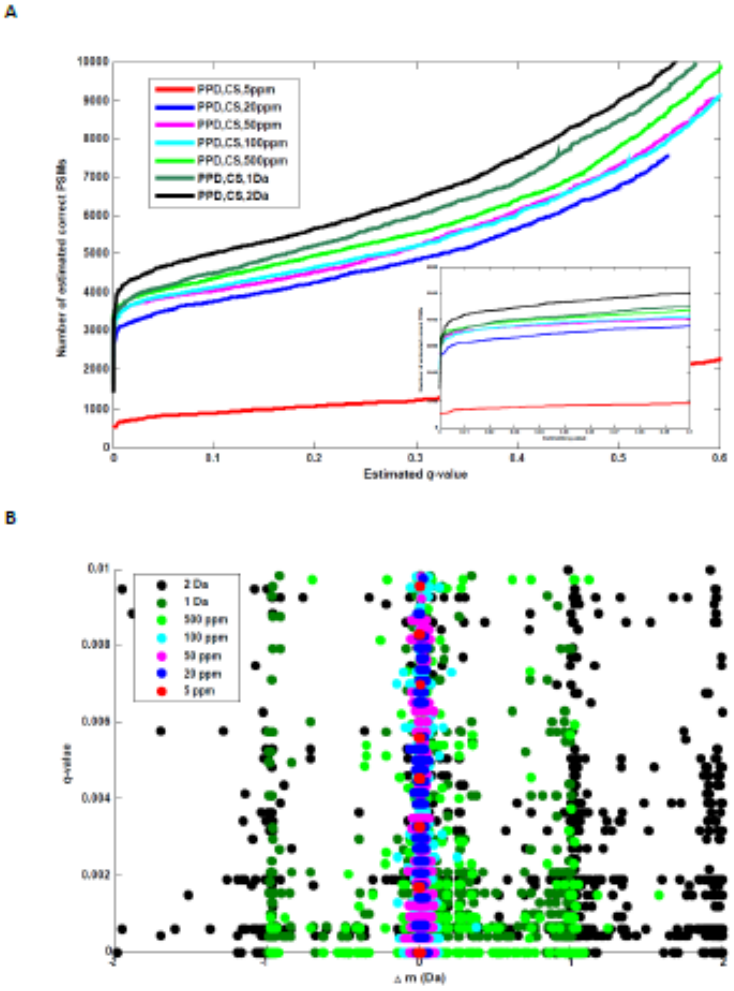


Figure 6. (A) Sensitivity comparison of peptide identification generated from seven parent mass tolerances. (B) The distribution of *q*-value with precursor mass error (in Daltons). PPD in CS mode was used.

- **The sensitivity improved with the wide precursor mass tolerance**. E.g. when *q*-value < 1%, when setting at 2 Da, 33% more peptides were obtained compared with 20 ppm, and about 5-fold more peptides were obtained compared with 5 ppm

- The mass error distribution indicated the main contributions to the peptide identifications might come from the spectra with miss-assigned monoisotopic masses

QC method: PPD
Search strategy: CS
Decoy design： REVTP

Conclusions

- For high mass accuracy data, when PPDistiller is applied
  - **Reversing tryptic peptide** (REVTP) is recommended for tryptic-digested sample data, because different decoy designs achieved similar or at least compatible sensitivity, but REVTP generated more accurate estimated *q*-values
  - **Composite search** is recommended, because it generated more accurate *q*-values without compromising the sensitivity
  - **Reasonable wide precursor mass tolerance** is recommended, because the sensitivity improved with wide precursor mass tolerances, and the precursor mass tolerance didn't affected the accuracy of estimated *q*-values

Acknowledgements

- Beijing Proteome Research Center
  - Jie Ma
  - Ning Li
  - Liwei Li
  - Dr. Songfeng Wu
  - Dr. Jianqi Li
  - Dr. Lei Dou
  - Prof. Fuchu He
  - Prof. Xiaohong Qian
- National University of Defense Technology
  - Prof. Hongwei Xie
  - Dr. Jiyang Zhang
- Ministry of Science & Technology, China
  - 973 program (2006CB910803, 2010CB912700)
  - 863 program (2006AA02A312)

Thanks for your attention!

zhuyp@hupo.org.cn