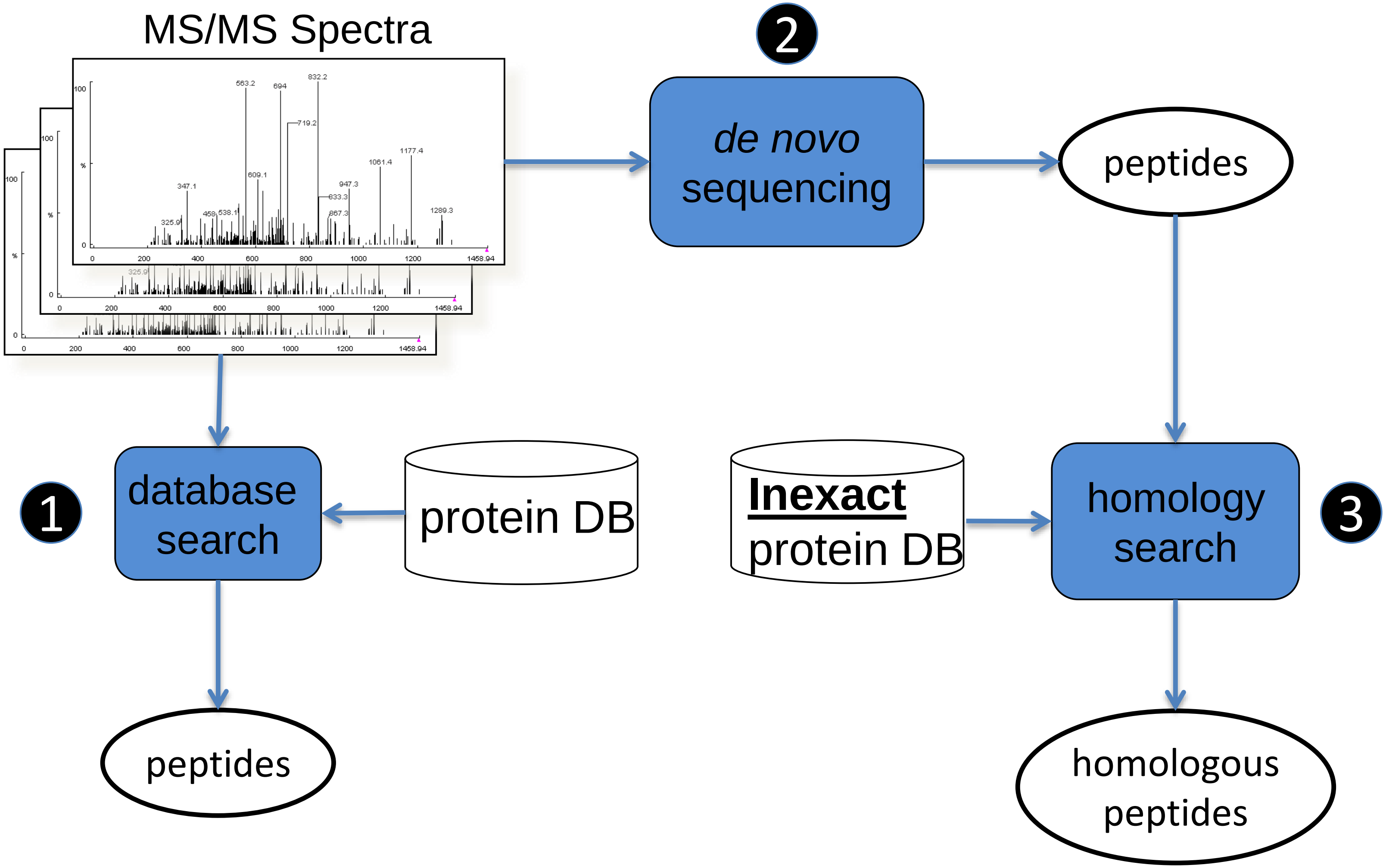
Integrating database search and de novo sequencing to improve the peptide identification

Baozhen Shan

Bioinformatics Solutions Inc.

1. Approaches for LC/MSMS data analysis



MS/MS Spectra

2 *de novo* sequencing → peptides → homology search

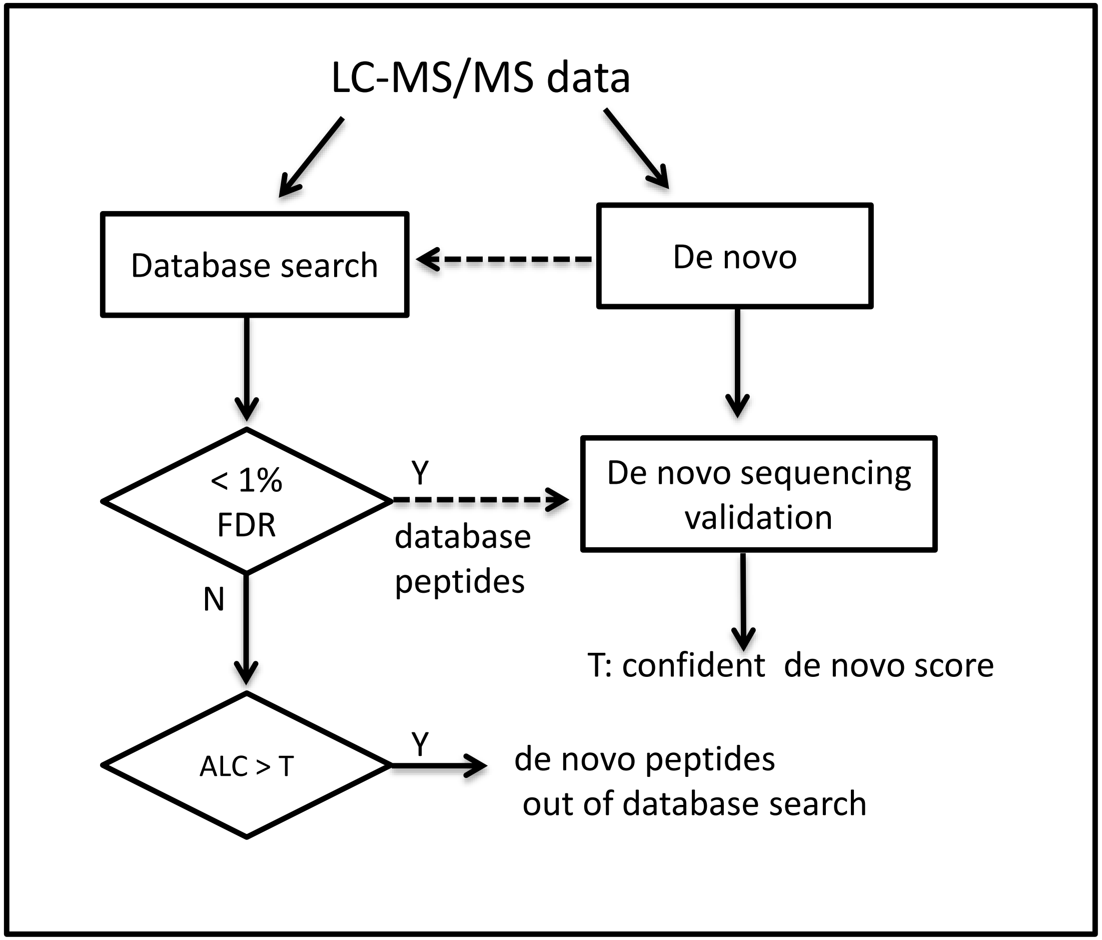
Inexact protein DB → homology search 3 → homologous peptides

1 database search ← protein DB

database search → peptides

Ma and Johnson. *MCP* (2012) 11: O111.014902

2. Integrating de novo and database search



LC-MS/MS data

Database search

De novo

< 1% FDR

Y

database peptides

De novo sequencing validation

N

T: confident de novo score

ALC > T

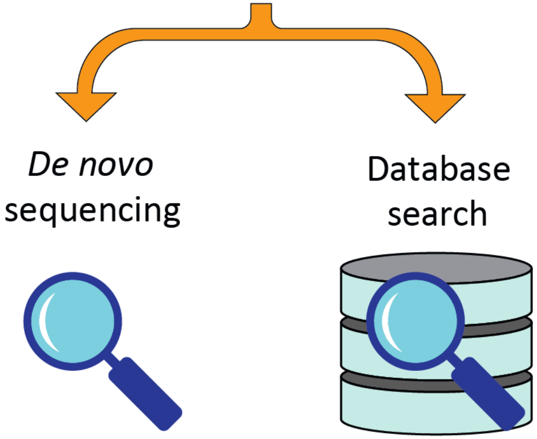
Y

de novo peptides out of database search

3. De novo sequencing improves DB search

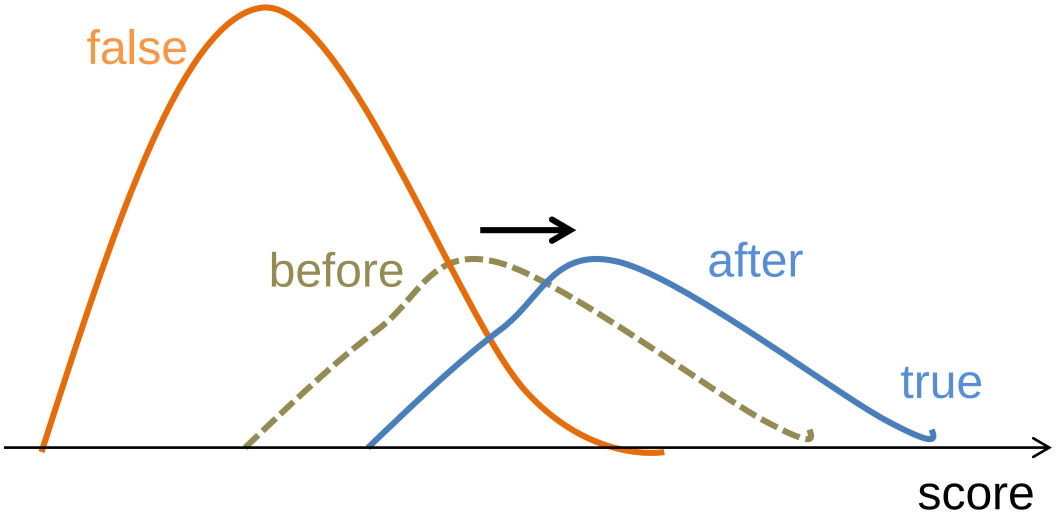
- Problem
  - Coverage
  - Modifications
  - Incomplete database

- Solution



3.1. Good scoring function

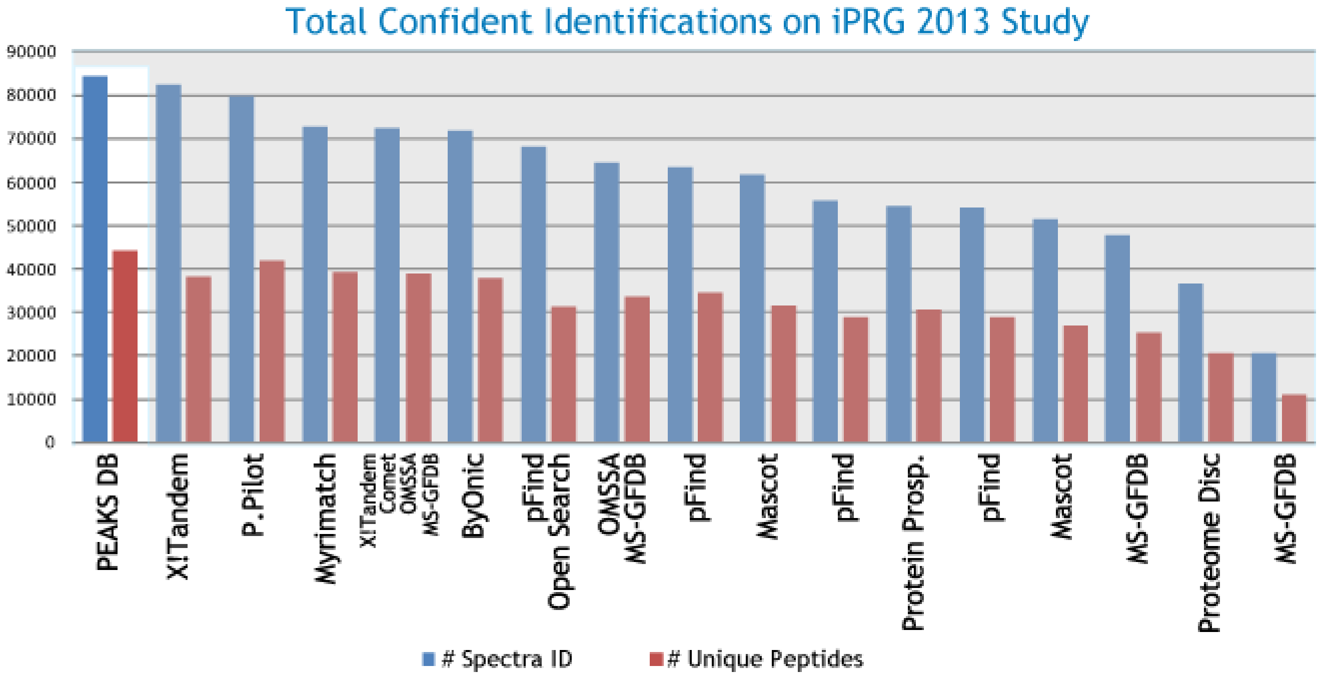
- Uses many more factors than other algorithms
  - particularly the similarity between *de novo* and DB sequence
  - many other scoring features considered
- Better separation of true and false means better accuracy and sensitivity.



Zhang et al, *MCP* (2012) 11, M111.010587.2011.

High sensitivity and accuracy

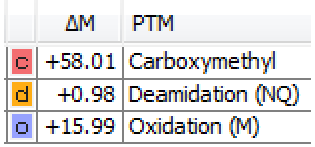
- ABRF iPRG 2013 study



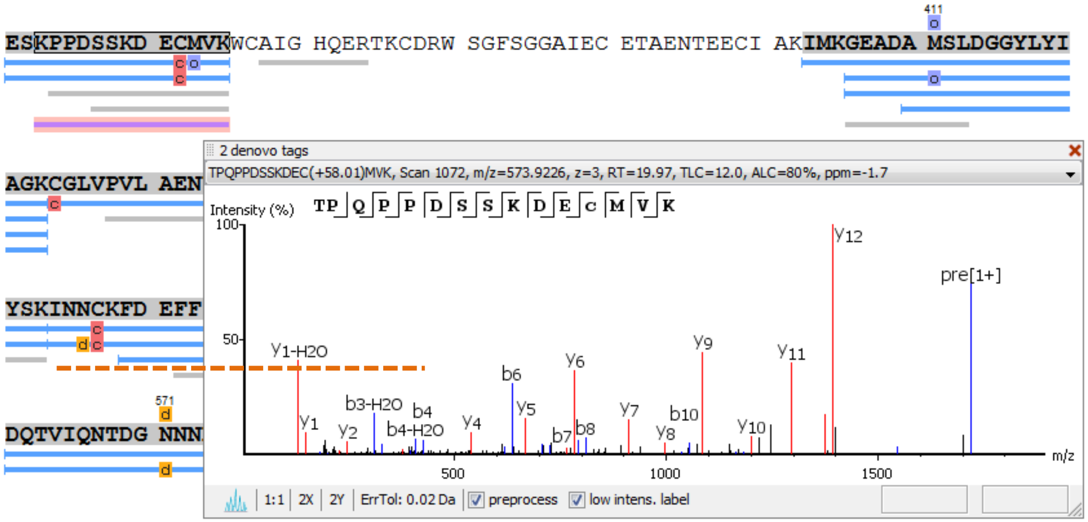
3.2. De novo assisted PTM “Blind Search”

- Search for PTM when there is a tag match.

PEAKS DB



	ΔM	PTM
c	+58.01	Carboxymethyl
d	+0.98	Deamidation (NQ)
o	+15.99	Oxidation (M)



ESKPPDSSKD ECMVKWCAIG HQERTKCDRW SGFSGGAIEC ETAENTEECI AKIMKGEADA MSLDGGYLYI

AGKCGLVPVL AEN

YSKINNCKFD EFF

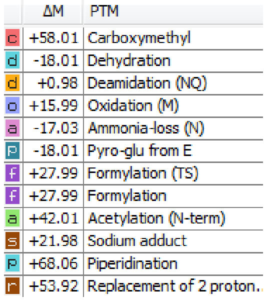
DQTVIQNTDG NNN

2 denovo tags

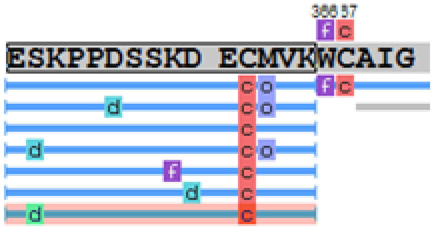
TPQPPDSSKDEC(+58.01)MVK, Scan 1072, m/z=573.9226, z=3, RT=19.97, TLC=12.0, ALC=80%, ppm=-1.7

1:1 | 2X | 2Y | ErrTol: 0.02 Da | ☑ preprocess ☑ low intens. label

PEAKS PTM



	ΔM	PTM
c	+58.01	Carboxymethyl
d	-18.01	Dehydration
d	+0.98	Deamidation (NQ)
o	+15.99	Oxidation (M)
a	-17.03	Ammonia-loss (N)
p	-18.01	Pyro-glu from E
f	+27.99	Formylation (TS)
f	+27.99	Formylation
a	+42.01	Acetylation (N-term)
s	+21.98	Sodium adduct
p	+68.06	Piperidination
r	+53.92	Replacement of 2 proton.



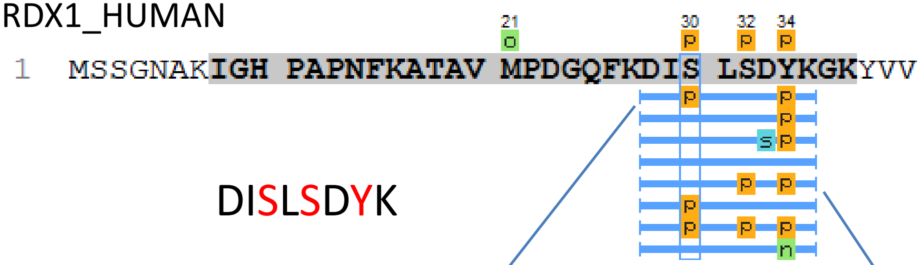
ESKPPDSSKD ECMVKWCAIG

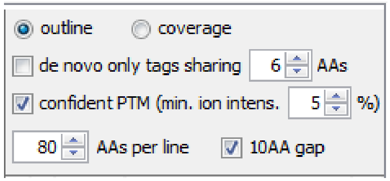
X. Han et al., *JPR* (2011), 10, 2930-2936.

ABRF iPRG 2012 study

One spiked peptide

PRDX1_HUMAN





● outline ○ coverage

□ de novo only tags sharing 6 AAs

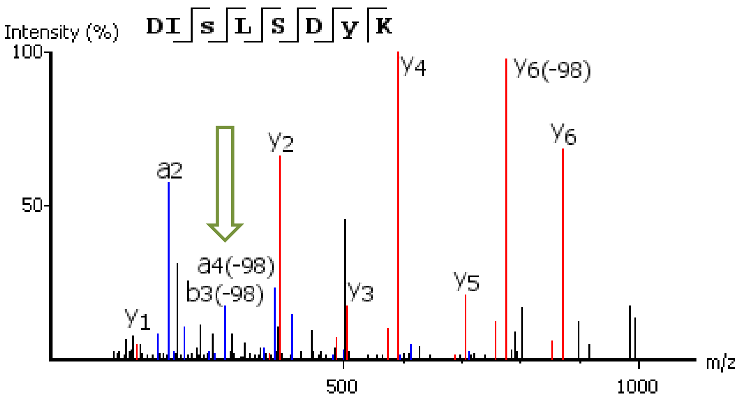
☑ confident PTM (min. ion intens. 5 %)

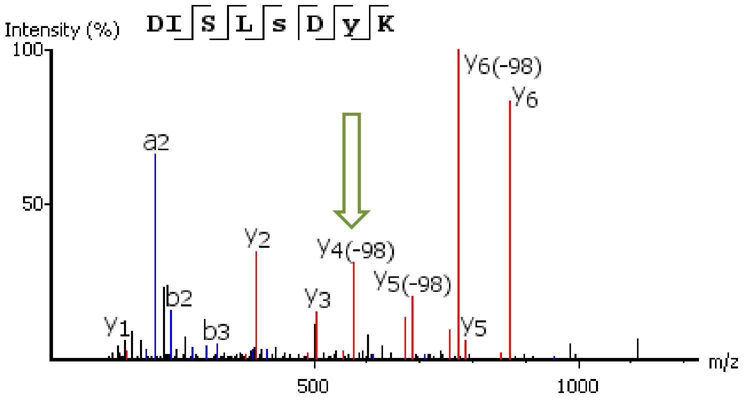
80 AAs per line ☑ 10AA gap

o	+15.99	Oxidation M
p	+79.97	Phosphorylation (STY)
s	+21.98	Sodium adduct
n	+44.99	Nitration Y

DISLSDYK

3,7-phospho *vs* 5,7-phospho

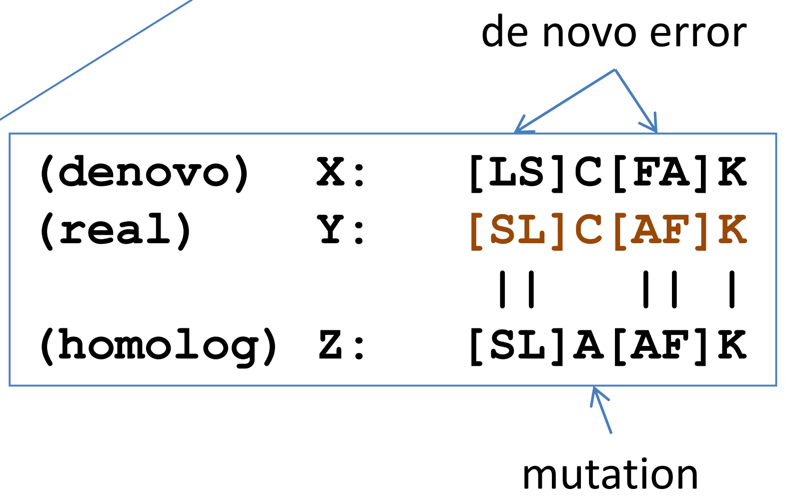




3.3. SPIDER homology search

Problem： de novo errors、database mutations

```
(denovo)   X:    LSCFAK
                      |
(homolog)  Z:    SLAAFK
```



de novo error

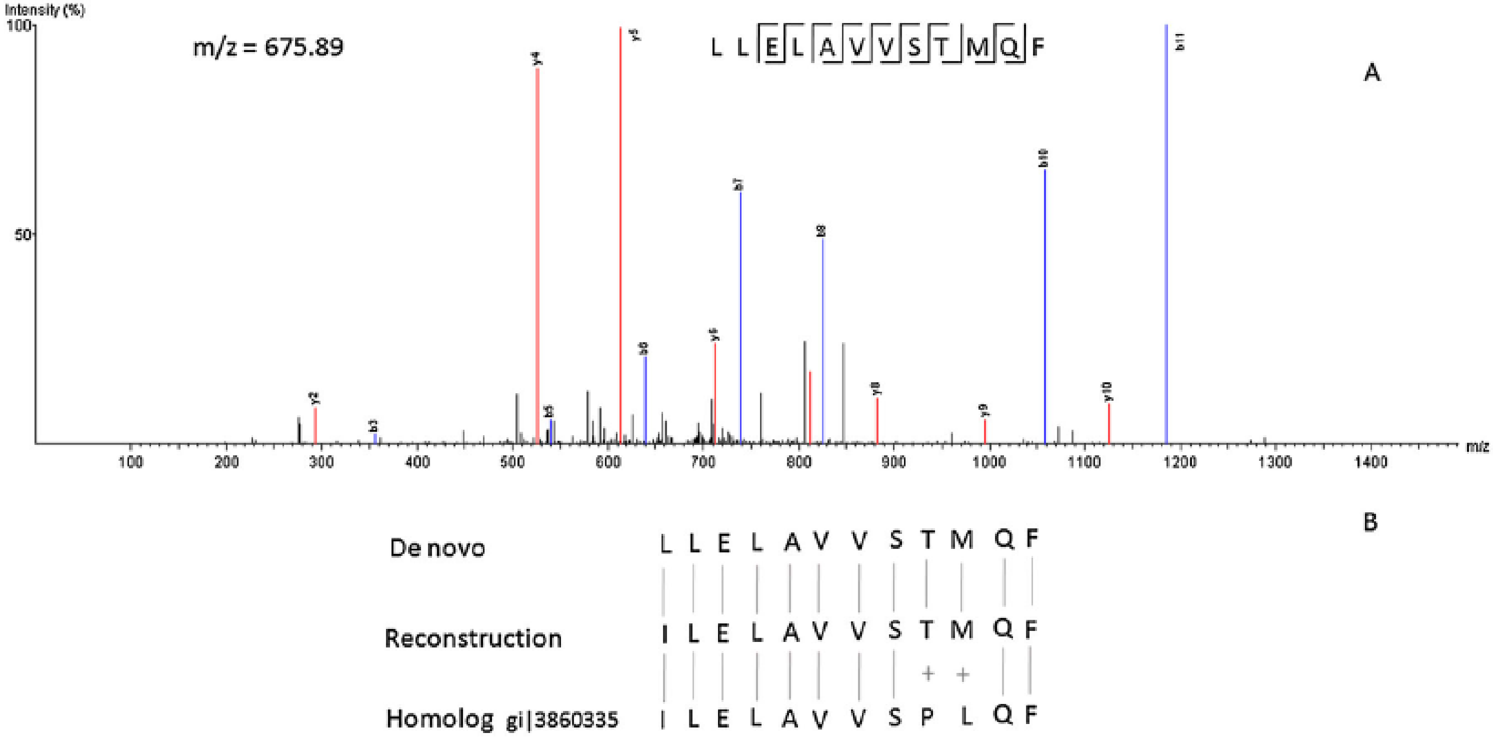
```
(denovo)   X:    [LS]C[FA]K
(real)     Y:    [SL]C[AF]K
                  ||   || |
(homolog)  Z:    [SL]A[AF]K
```

mutation

Solution： minimize de novo errors and mutations

Y. Han *et al*, *JBCB* (2005) 3, 697-716.

Peptides in camel milk



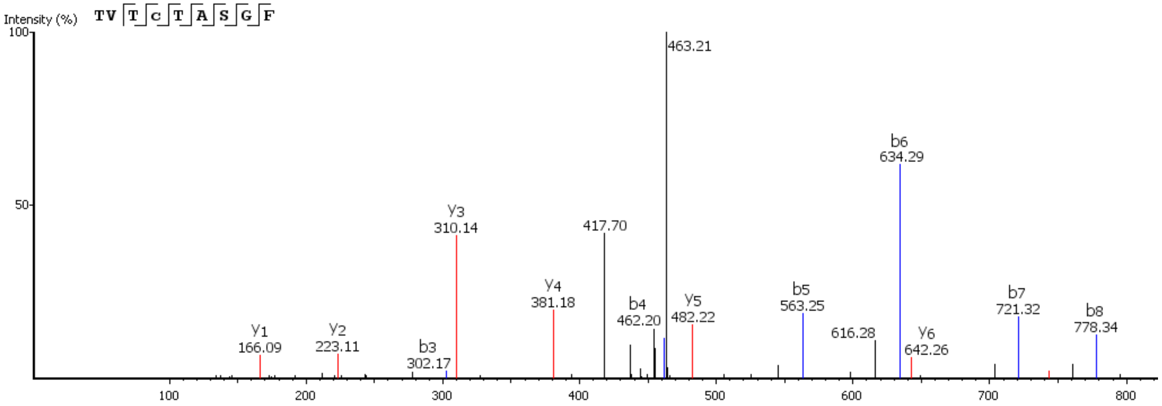
J. Mass Spectrom. **2013**, *48*, 779–794

4. DB search validates *De novo* sequencing

- Problem

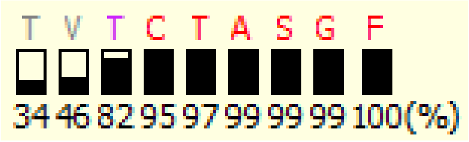
  De novo sequencing

  - Ambiguity of de novo sequence
  - Partially correct tags

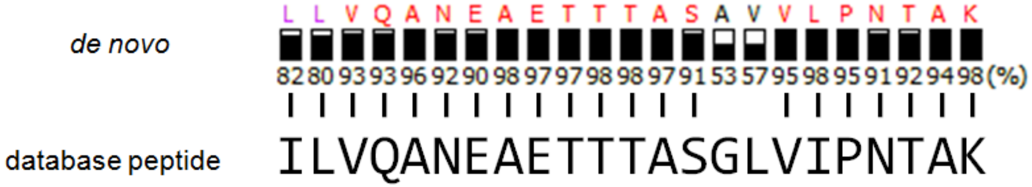


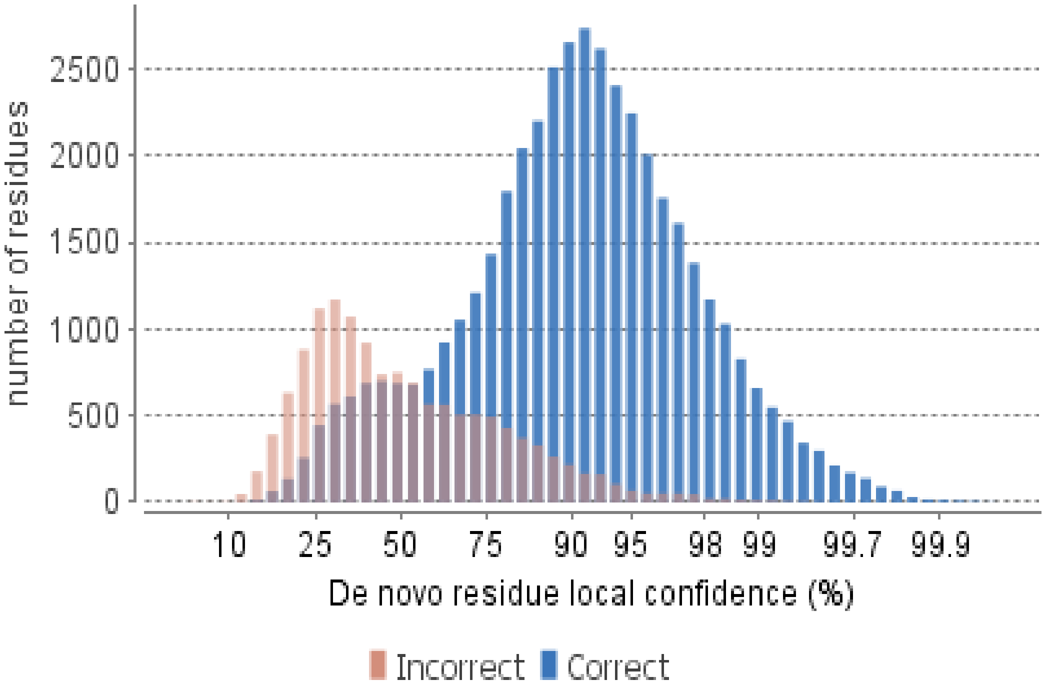
- Solution

  Local confidence score

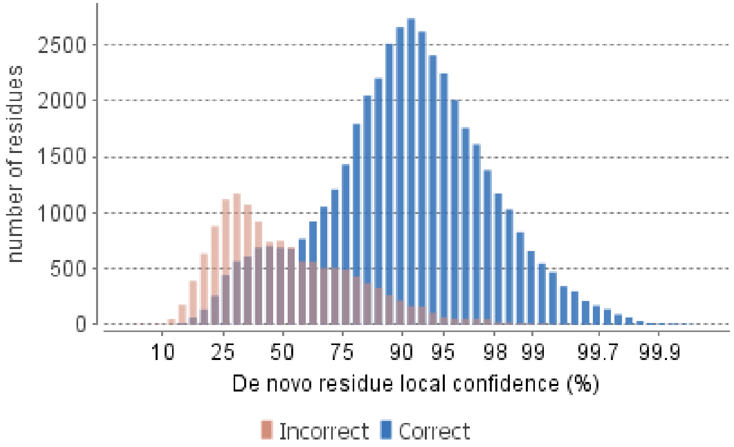


Validation with DB peptides

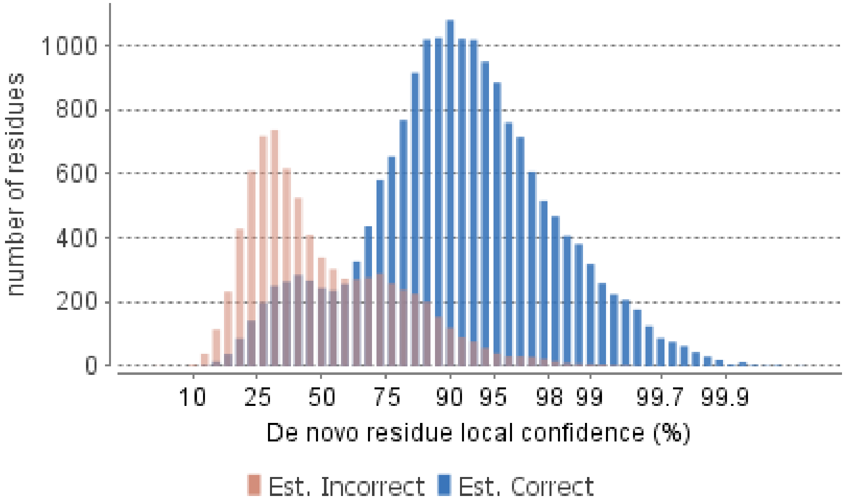




De novo - only peptides



de novo peptides validated by DB



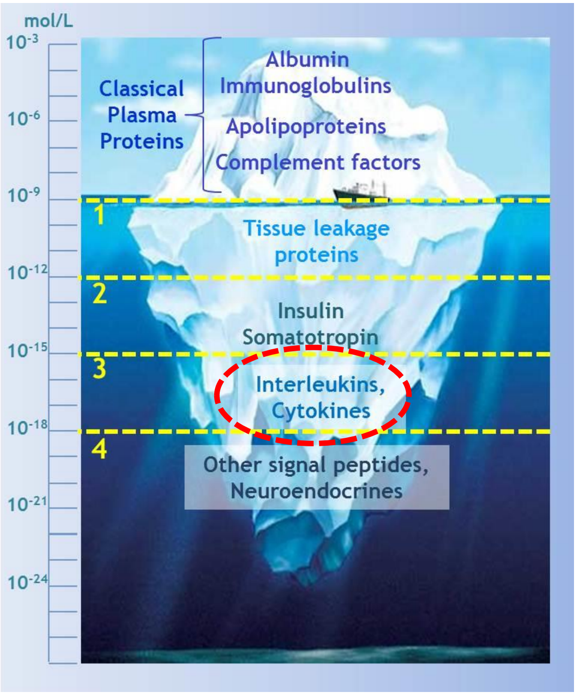
score distribution
of *de novo* “only” peptides
with estimated correctness

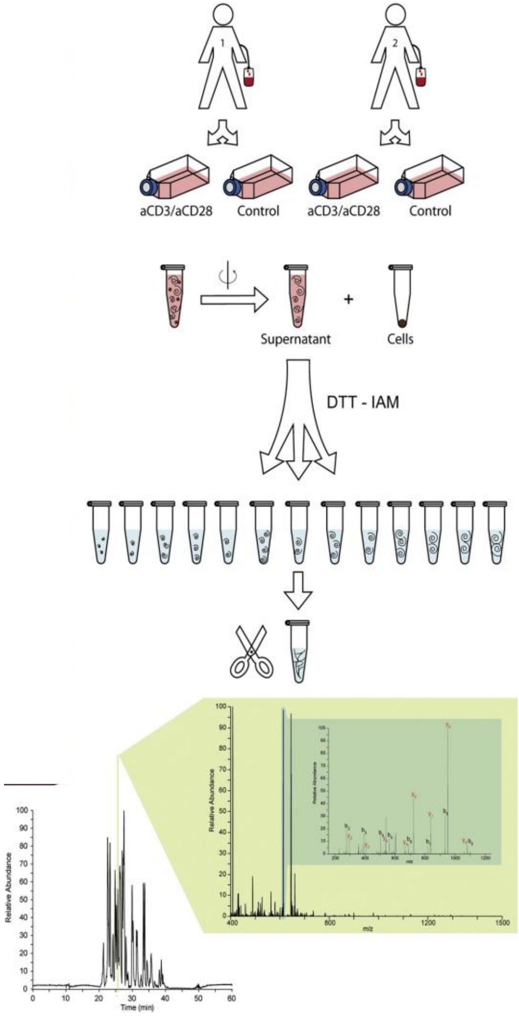
5. Finding endogenous peptides

- Extracellular proteome

  low abundance, esp. signaling peptides

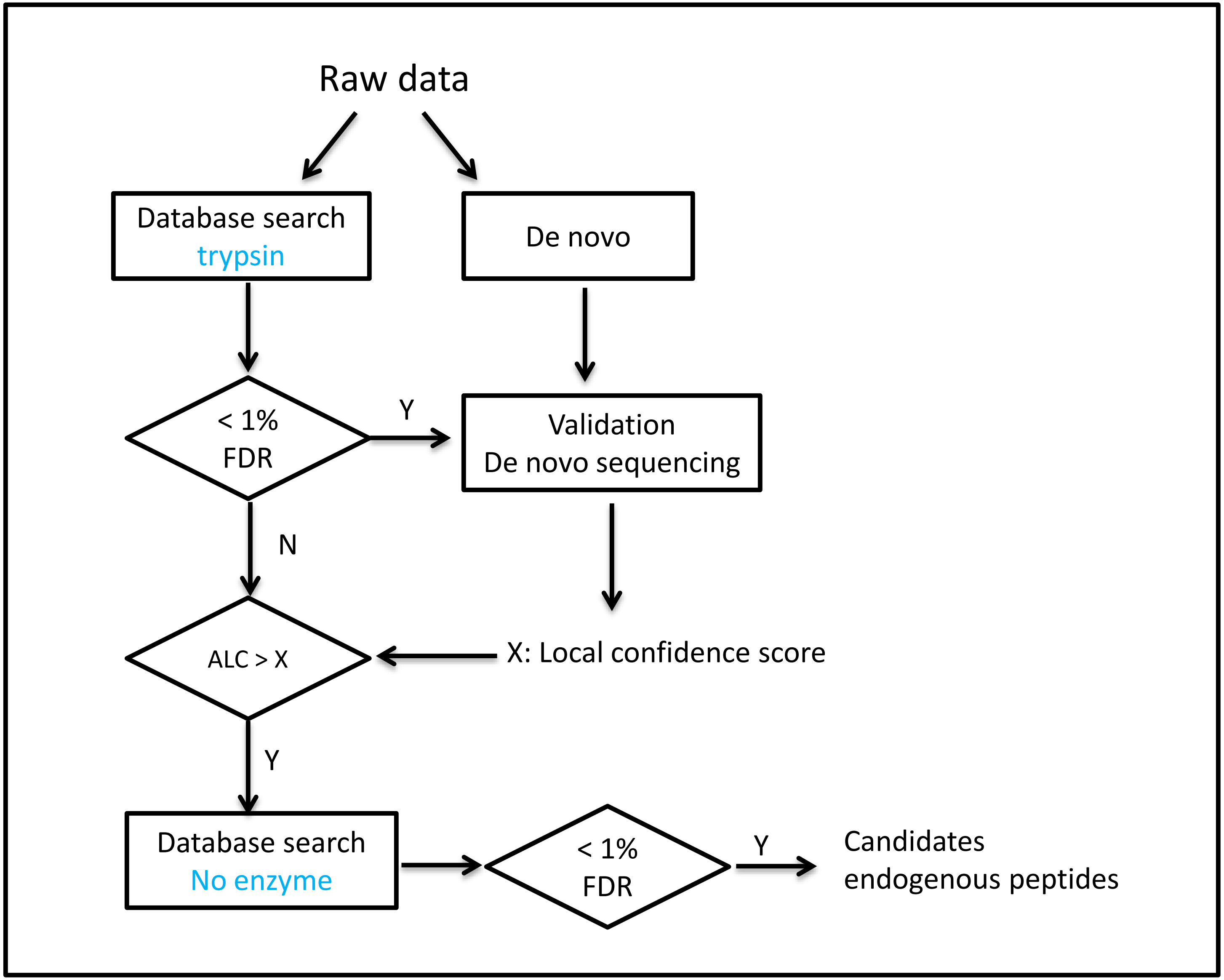
  contaminated by intracellular proteins





Finoulst et al. J. Proteomics (2012)

Workflow for endogenous peptides identification



Raw data

Database search
trypsin

De novo

< 1%
FDR

Y

Validation
De novo sequencing

N

ALC > X

X: Local confidence score

Y

Database search
No enzyme

< 1%
FDR

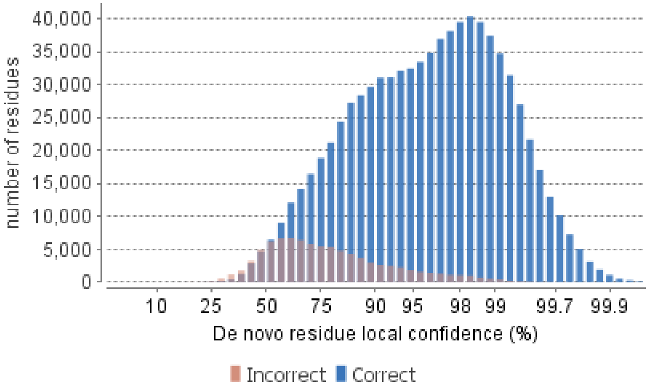
Y

Candidates
endogenous peptides

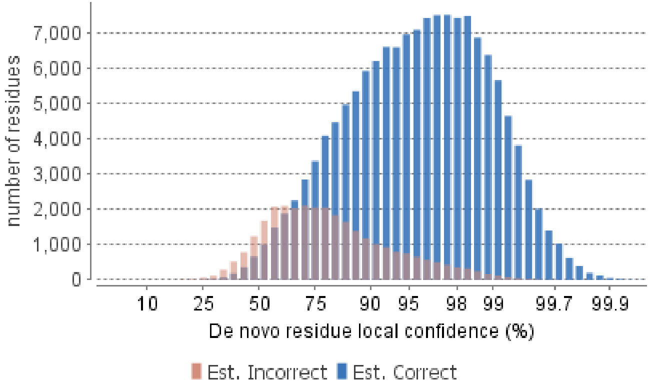
Identification of peptides

	# MS/MS	# peptides
LC-MS/MS	1954303	
Database search	584614	18625
De novo sequencing	15597	987

validated

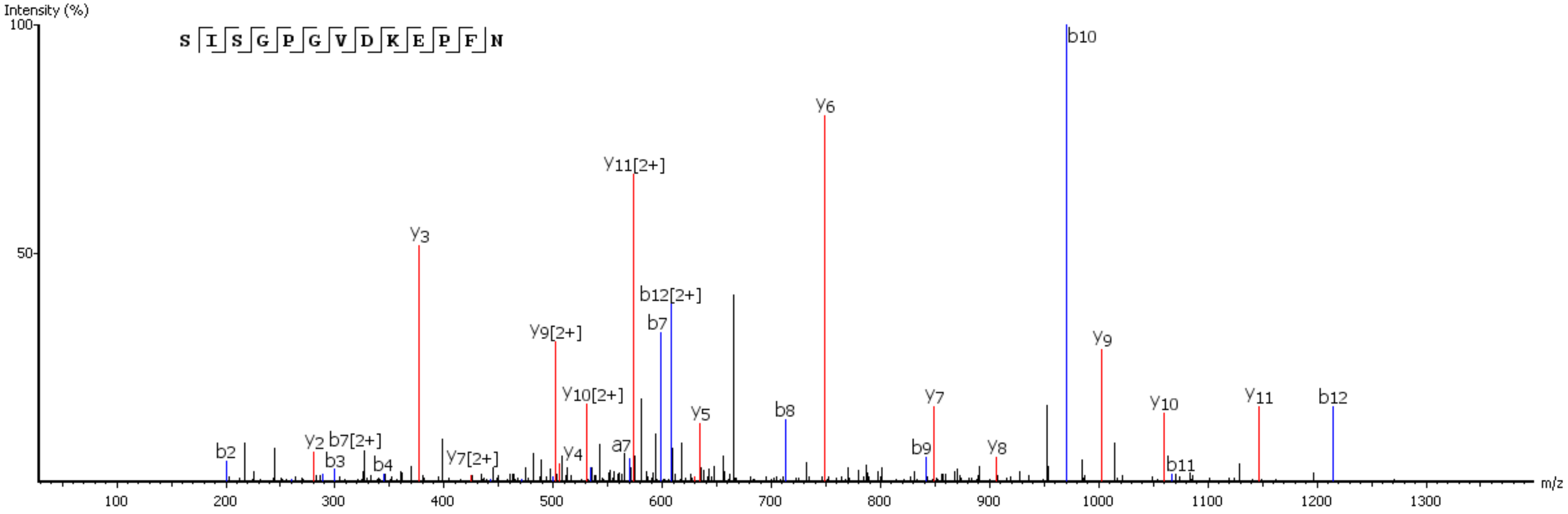


estimated



70 Human non-tryptic peptides

Example of an endogenous peptide

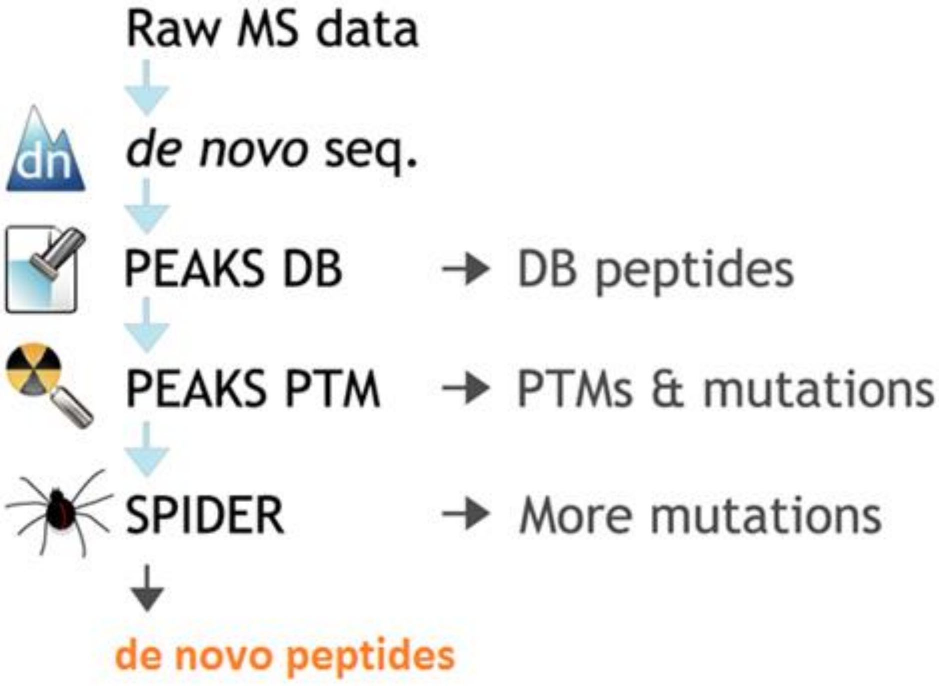


Q08554|DSC1_HUMAN

a member of the desmocollin subfamily

extracellular region

Implement in PEAKS



Acknowledgement

- PEAKS R&D team at BSI
- Collaborations

  Prof. Bin Ma at UW

  Prof. Gilles A. Lajoie at UWO

  Prof. Kaizhong Zhang at UWO

  Prof. Peter Verhaert at Delft University of Technology

  Cuijie Zhang at Samuel Lunenfeld Research Institute