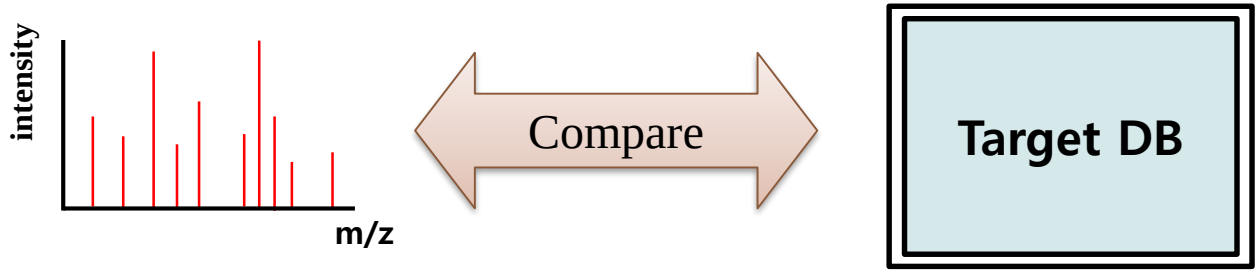
Target-Small Decoy Search Strategy for False Discovery Rate

2020.06

김현우
한국과학기술정보연구원

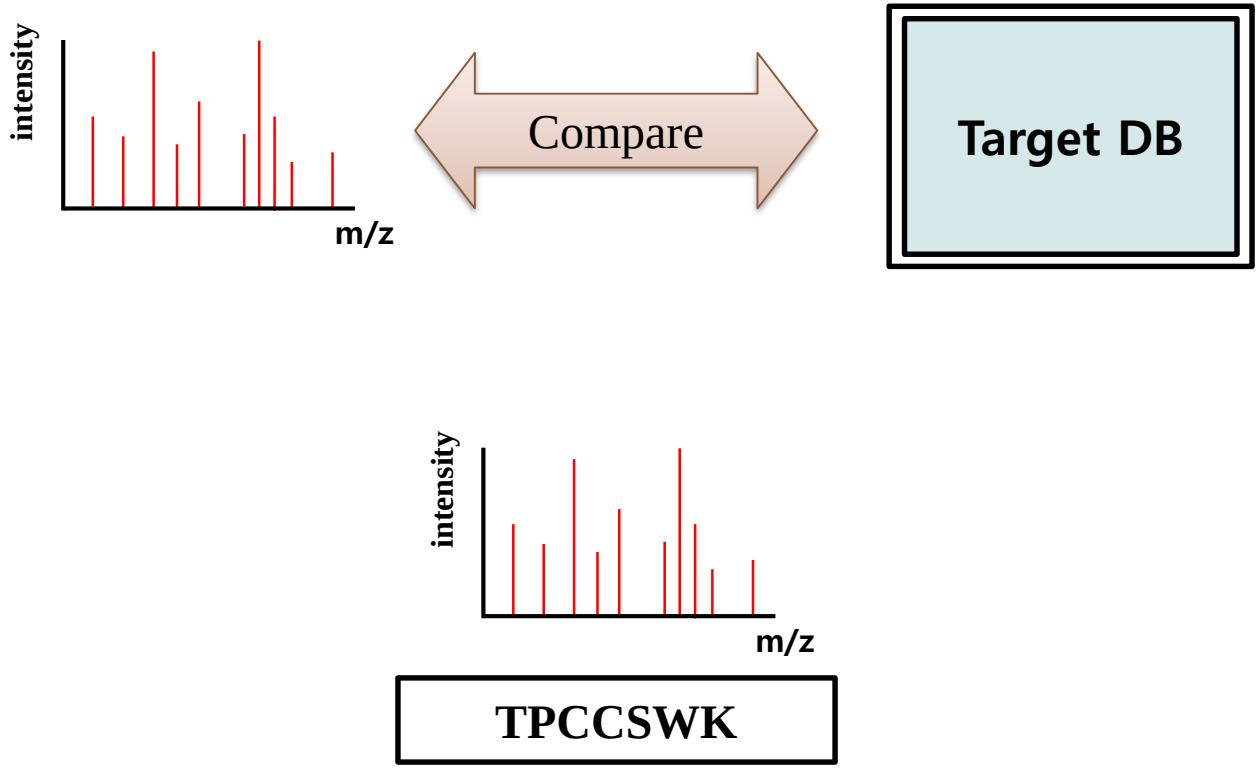
INTORDUCTION

- **Peptide identification**



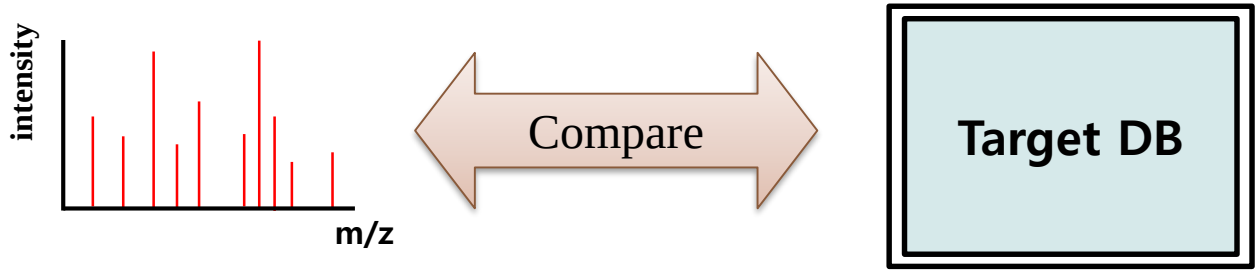
INTORDUCTION

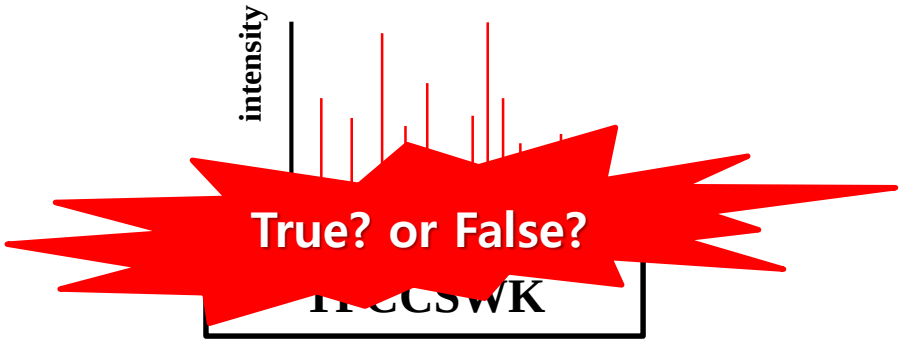
- **Peptide identification**



INTORDUCTION

- **Peptide identification**





INTORDUCTION

- **False Discovery Rate (FDR)**
  - 1% FDR

INTORDUCTION

- **False Discovery Rate (FDR)**
  - 1% FDR

1	**LNRSDHFHSR**
2	**LDMSFHSR**
⋮	
***i* – 1**	**TPCCSWK**
i	**DHGIFHSR**
***i* + 1**	**MGIFHSR**
⋮	
***n* – 1**	**SRCHSHK**
n	**YEYEVDKDFSSK**

INTORDUCTION

- **False Discovery Rate (FDR)**
  - 1% FDR

$$\frac{1}{100} \times n \text{ 개 FALSE!}$$

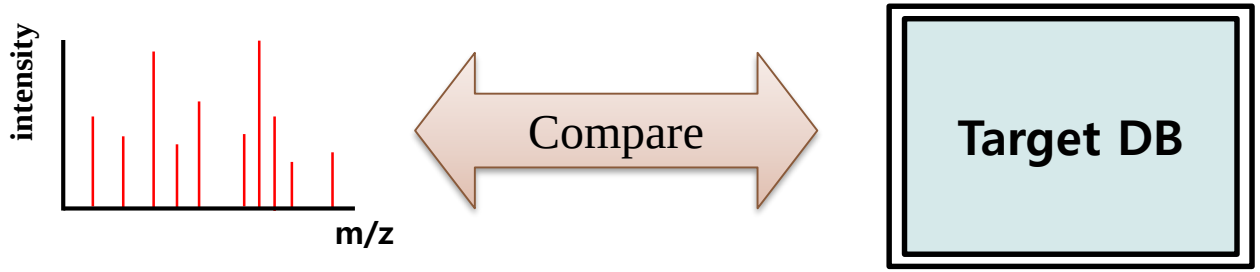
1	LNRSDHFHSR
2	LDMSFHSR
⋮	
$i-1$	TPCCSWK
i	DHGIFHSR
$i+1$	MGIFHSR
⋮	
$n-1$	SRCHSHK
n	YEYEVDKDFSSK

INTORDUCTION

- **FDR estimation**
  - Target-decoy search strategy

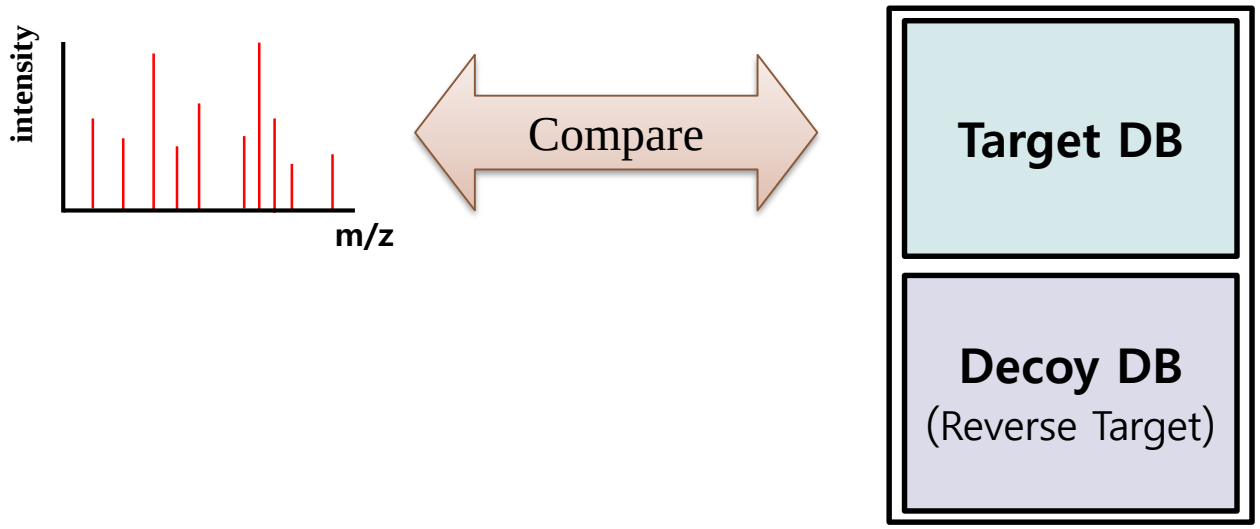
INTORDUCTION

- **Peptide identification**



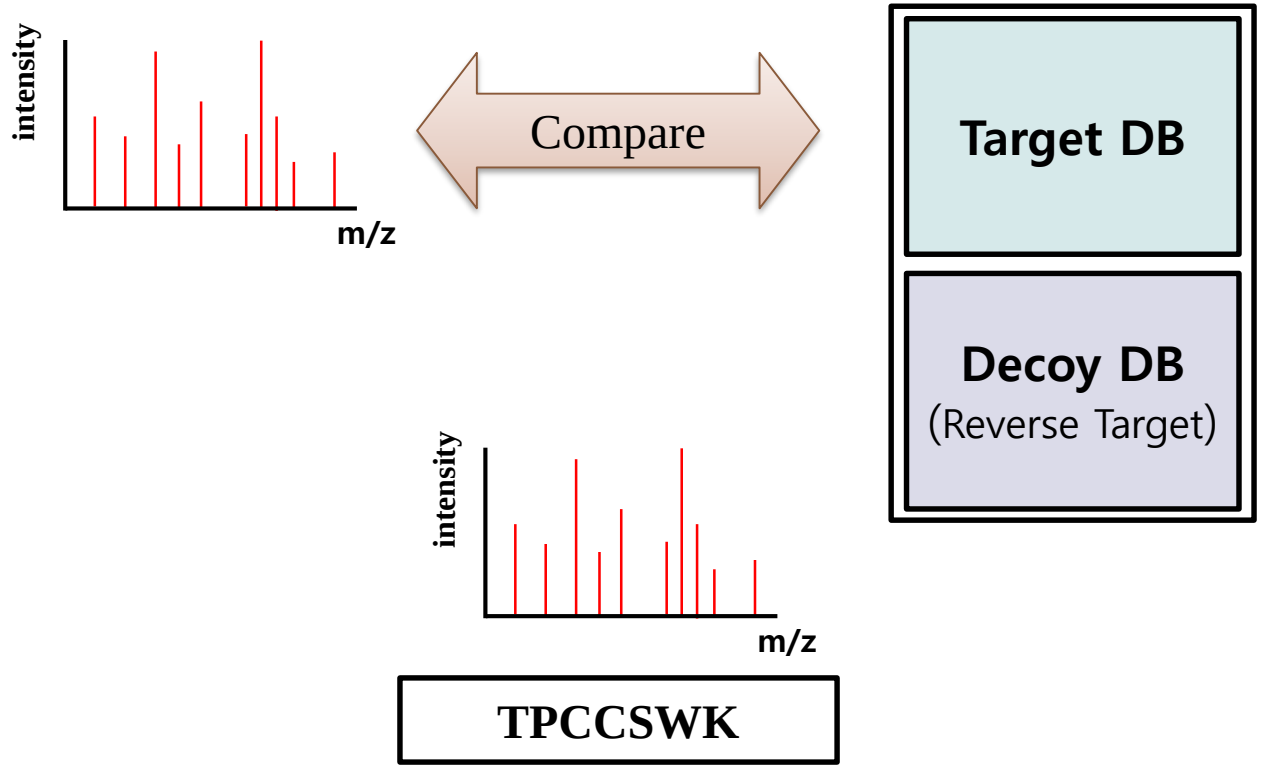
INTORDUCTION

- **Peptide identification**



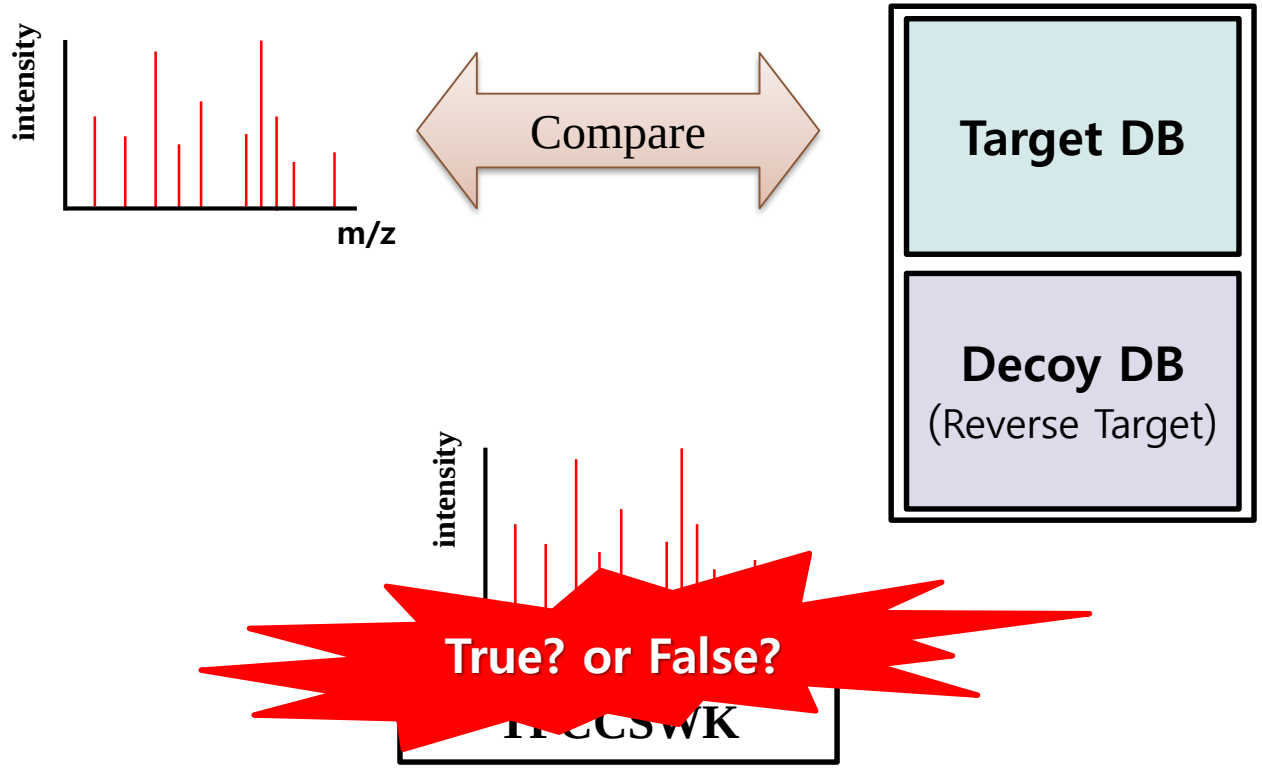
INTORDUCTION

- **Peptide identification**



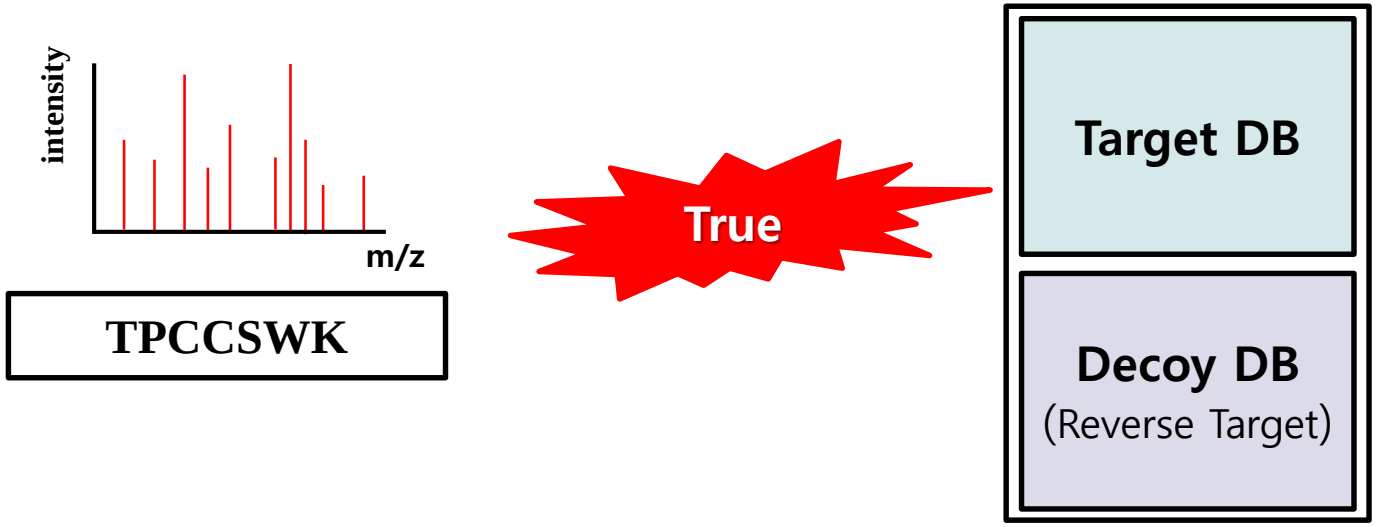
INTORDUCTION

- **Peptide identification**



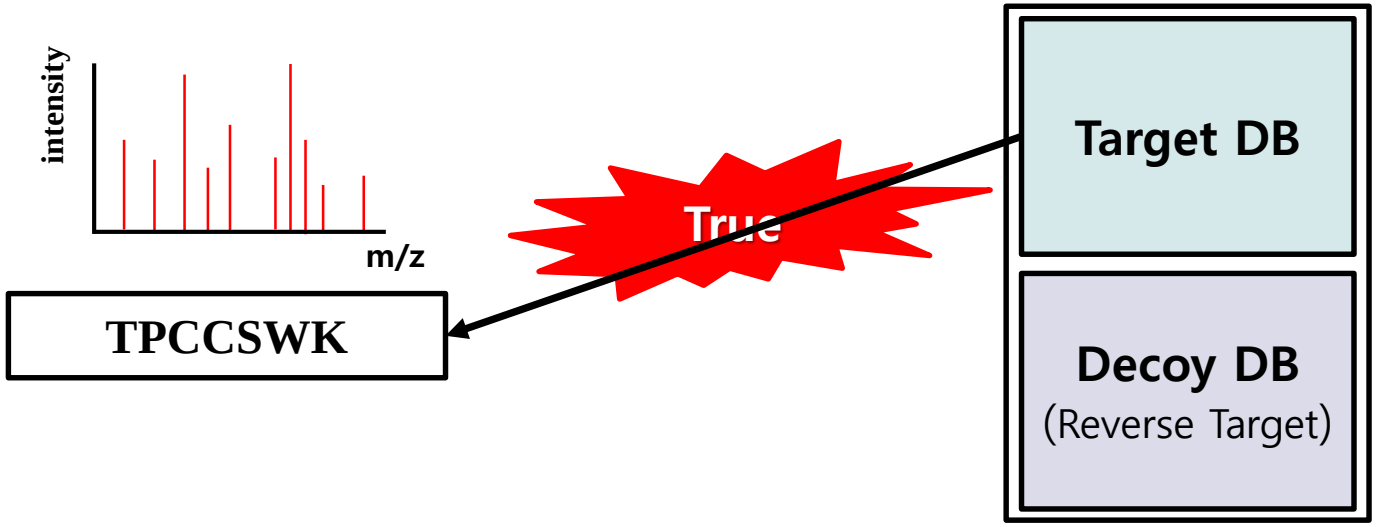
INTORDUCTION

- **Peptide identification**



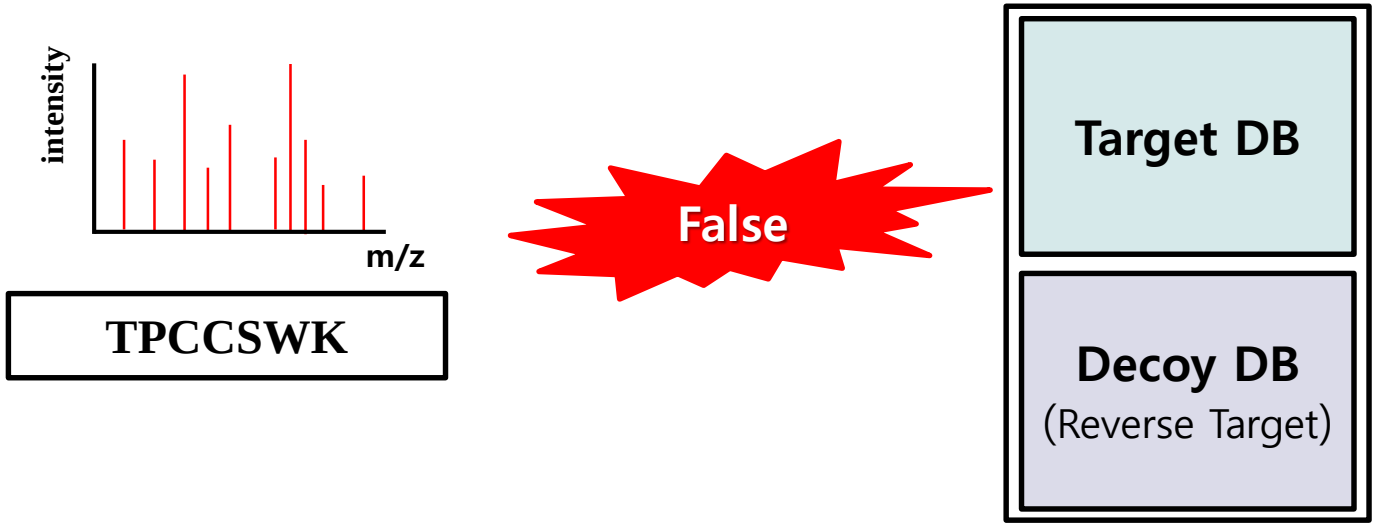
INTORDUCTION

- **Peptide identification**



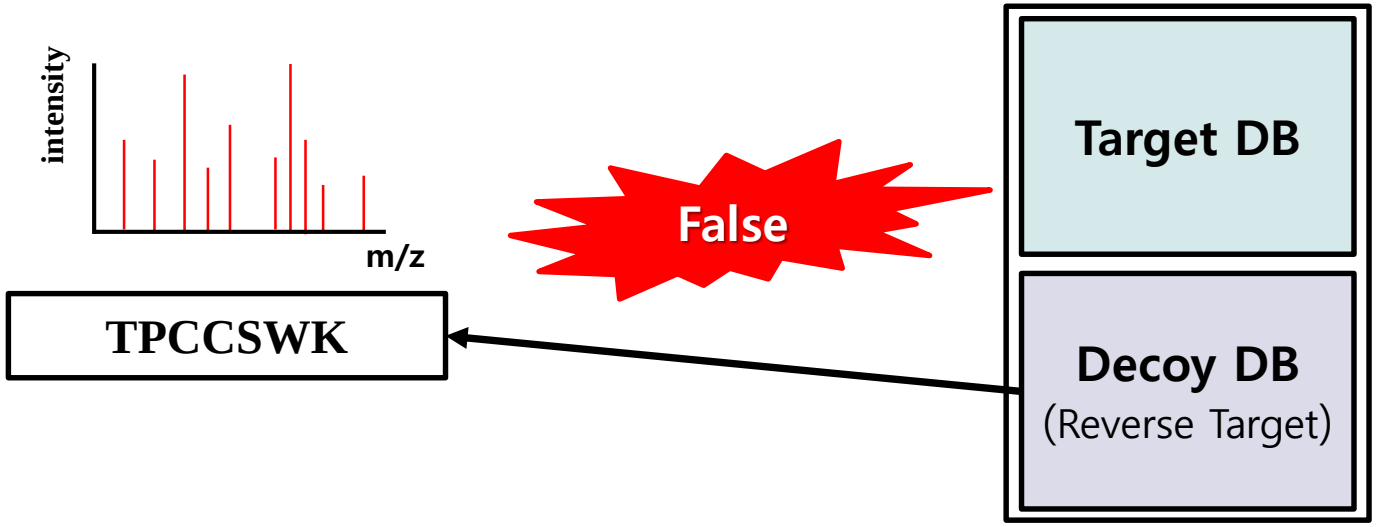
INTORDUCTION

- **Peptide identification**



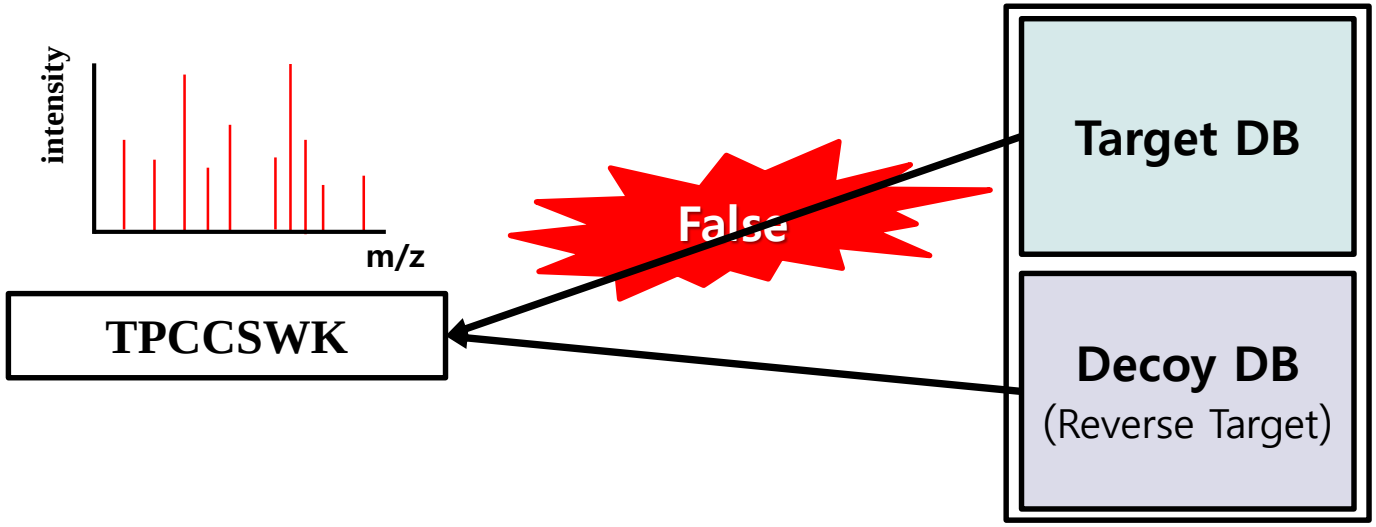
INTORDUCTION

- **Peptide identification**



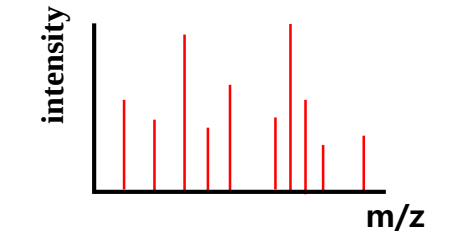
INTORDUCTION

- **Peptide identification**



INTORDUCTION

- **Peptide identification**



TPCCSWK

False

Target DB

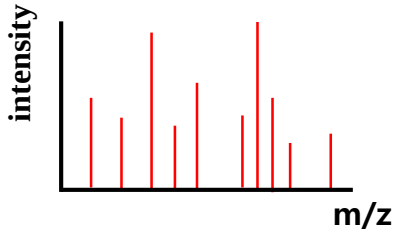
Decoy DB
(Reverse Target)

Target ?

Decoy ?

INTORDUCTION

- **Peptide identification**



TPCCSWK

False

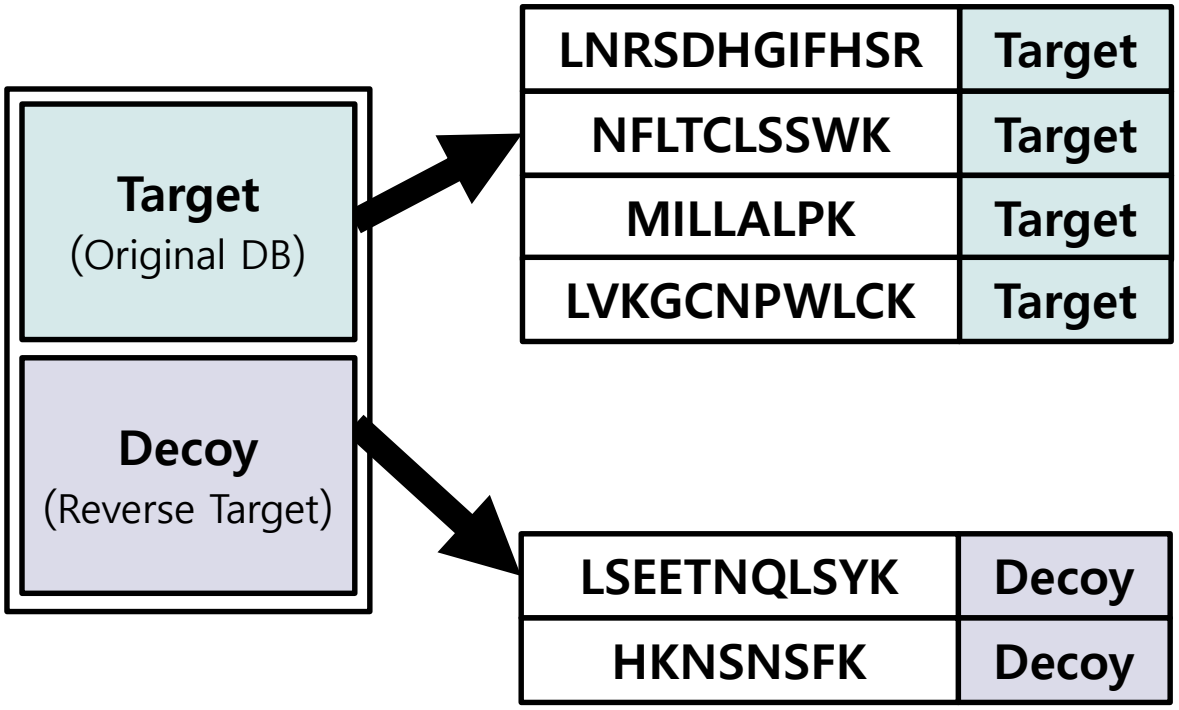
Target DB

Decoy DB
(Reverse Target)

Target ?	Decoy ?
50%	50%

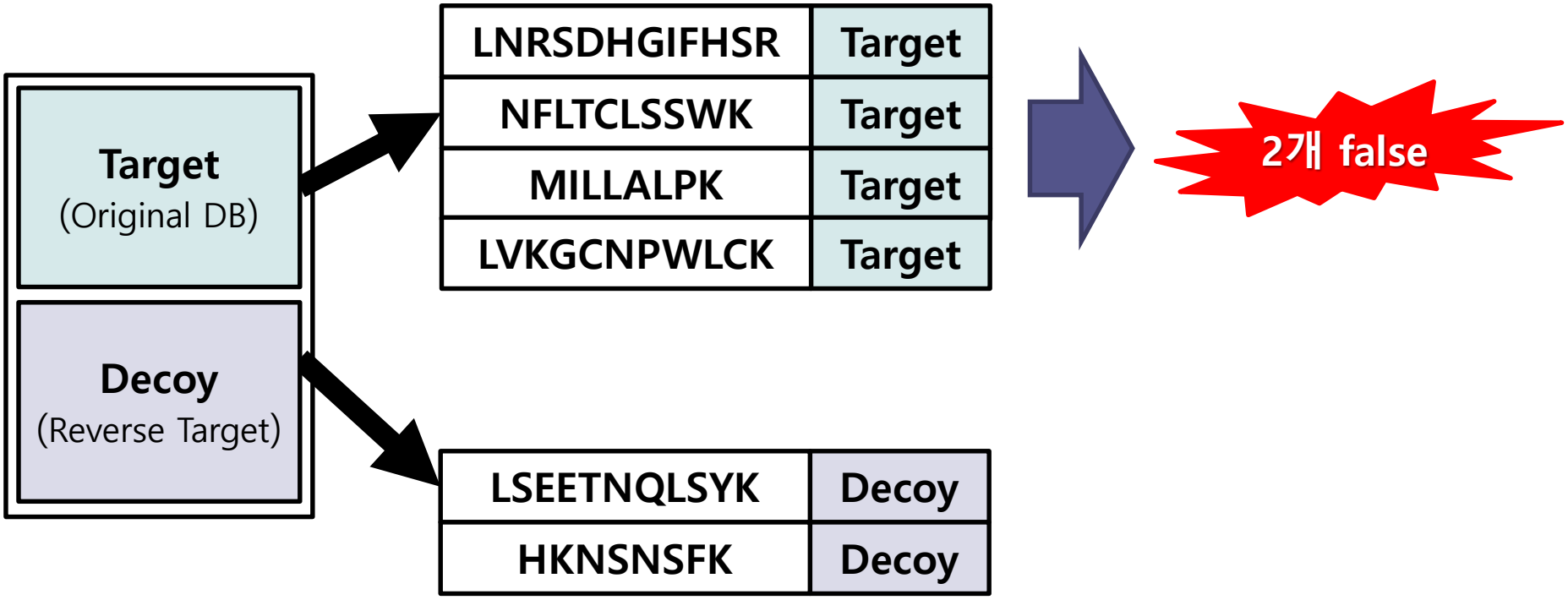
INTORDUCTION

- **FDR estimation**



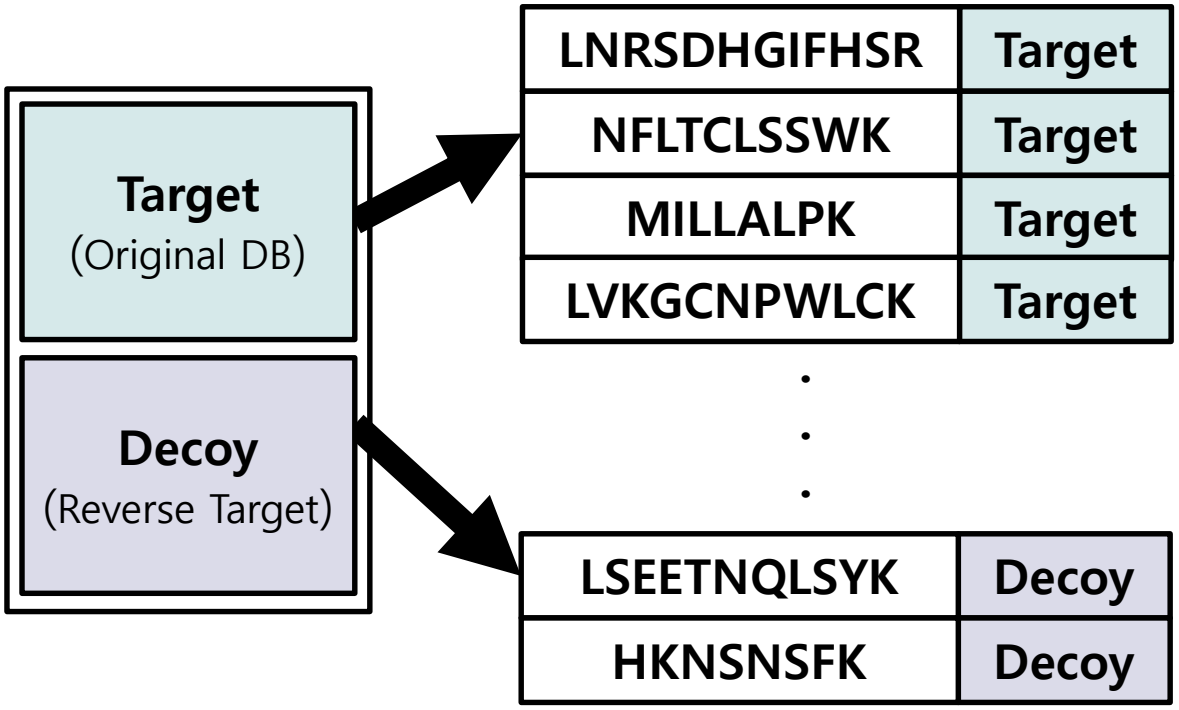
INTORDUCTION

- **FDR estimation**



INTORDUCTION

- **FDR estimation**



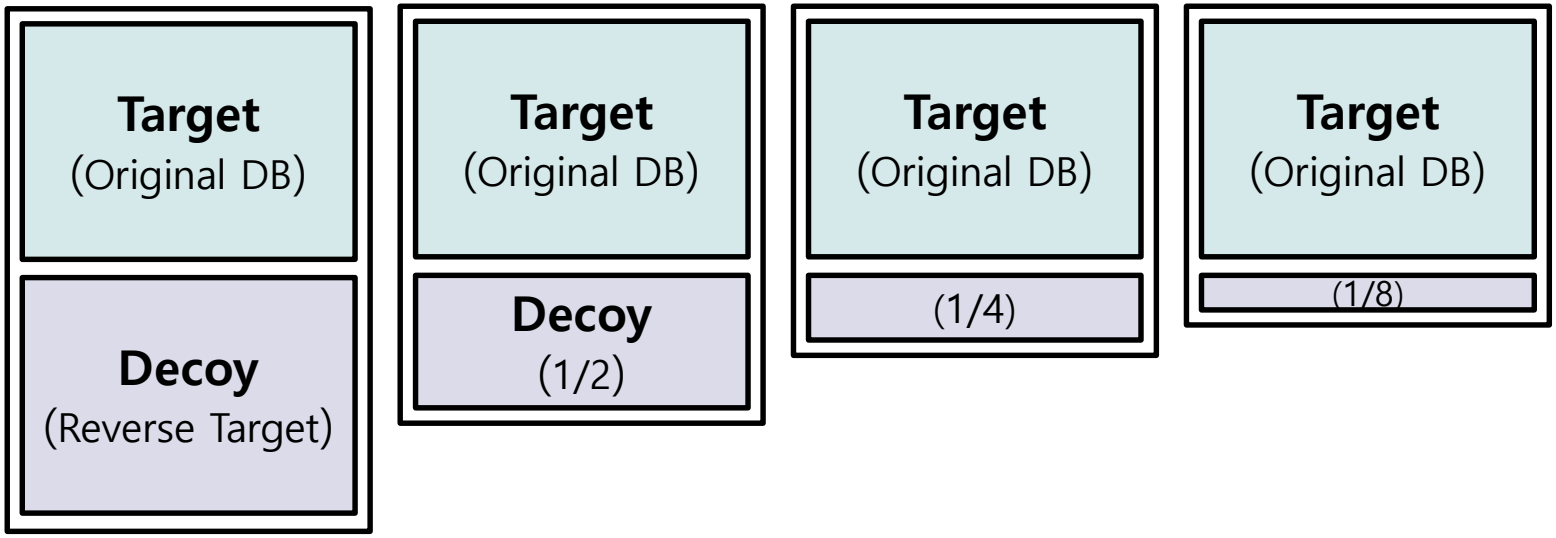
$$\frac{\#Decoy}{\#Target} = FDR$$

INTORDUCTION

- **Target-decoy search strategy**
  - Increases running time to search peptides against a protein database
    - The database size is twice as large as original protein database size
  - Have an important problem of sensitivity

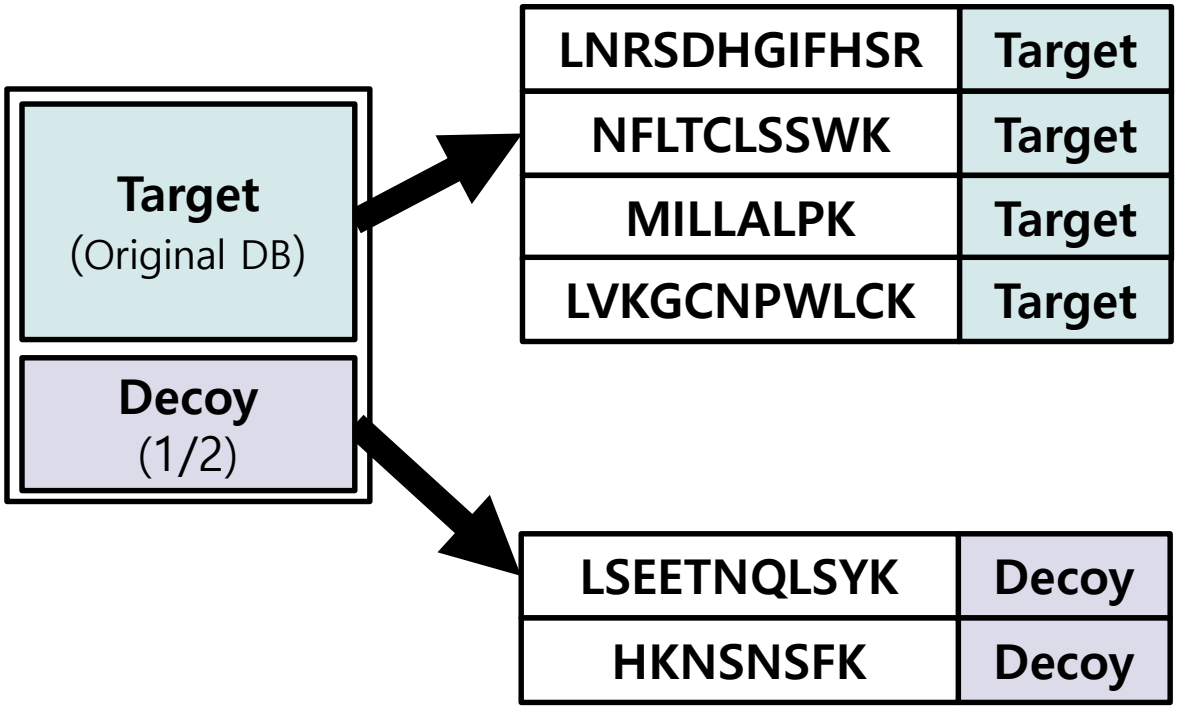
METHOD

- **Small Decoy database**
  - Decoy sequences are generated by reversing target sequences
  - Small decoy sequences are randomly selected in the decoy sequences



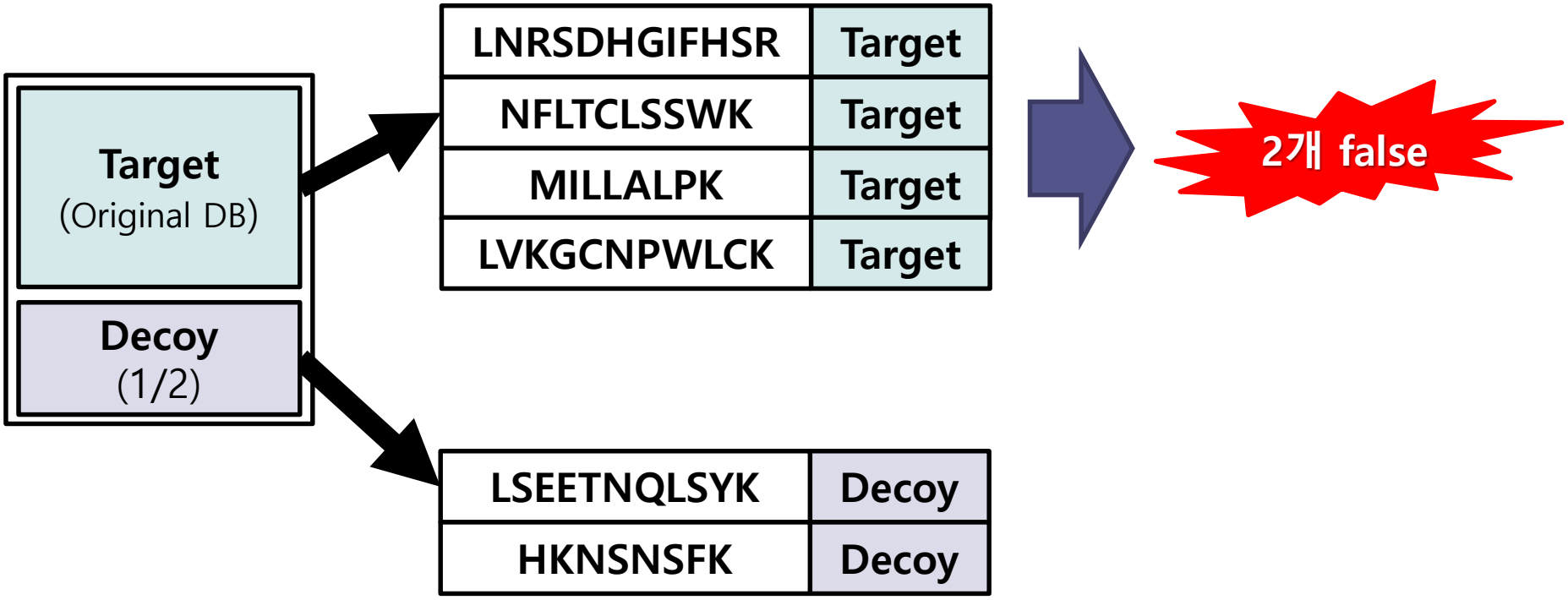
METHOD

- **FDR estimation using small decoy database**



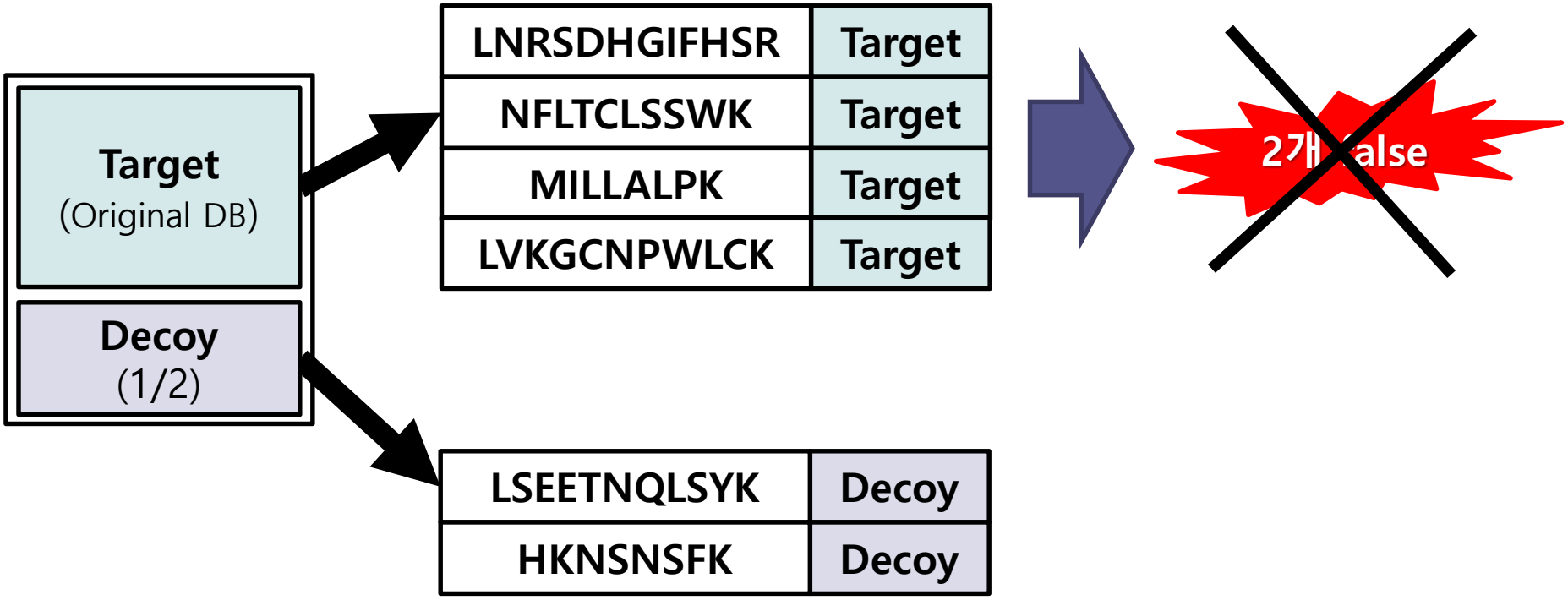
METHOD

- **FDR estimation using small decoy database**



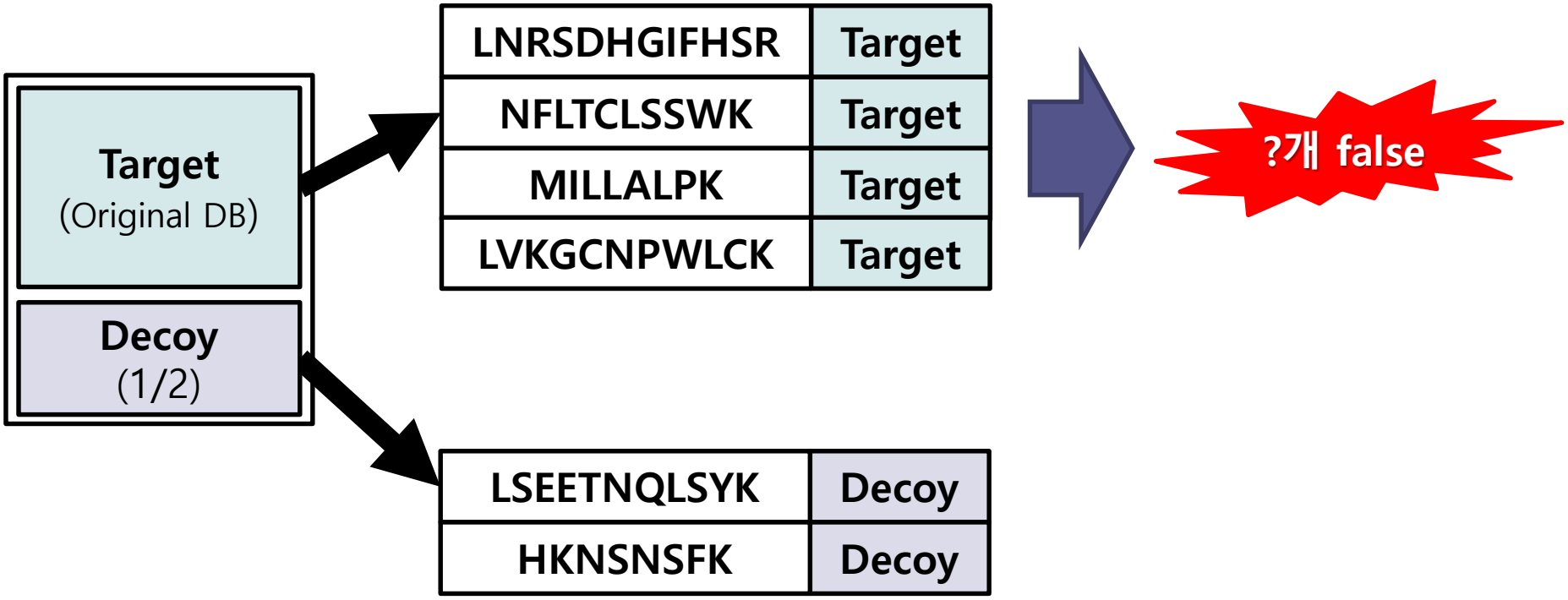
METHOD

- **FDR estimation using small decoy database**



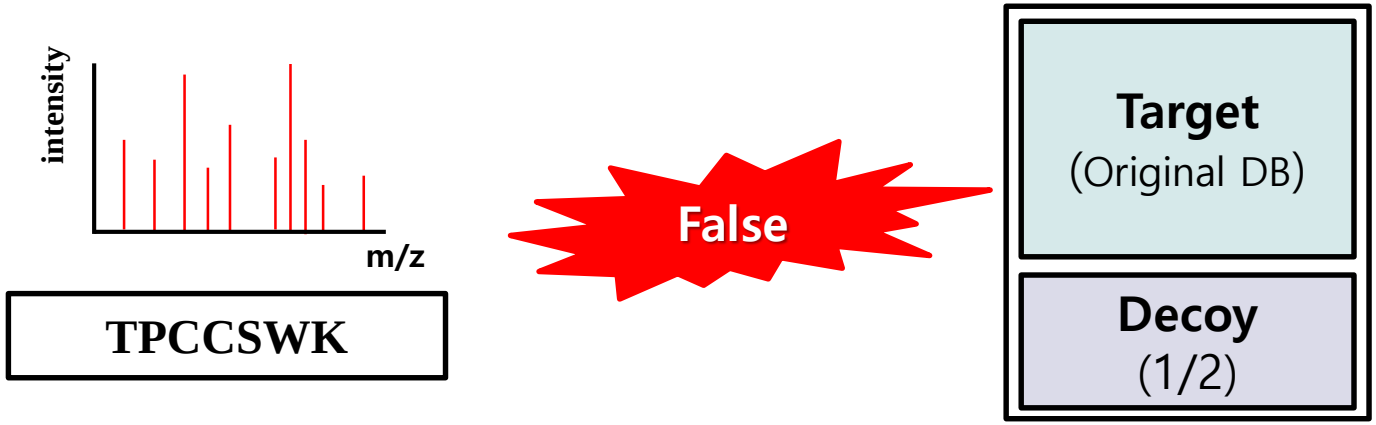
METHOD

- **FDR estimation using small decoy database**



METHOD

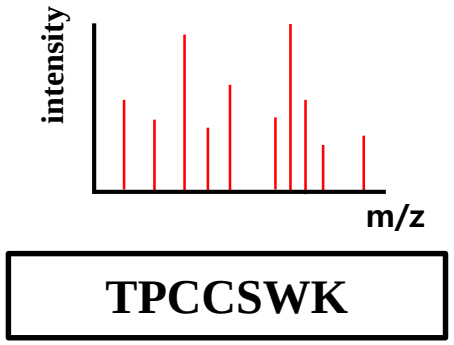
- **Peptide identification using small decoy database**

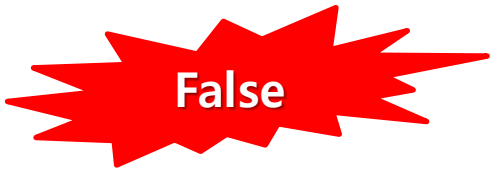


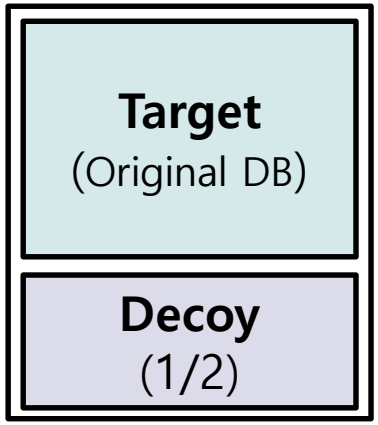
Target ?	Decoy ?
??%	??%

METHOD

- **Peptide identification using small decoy database**



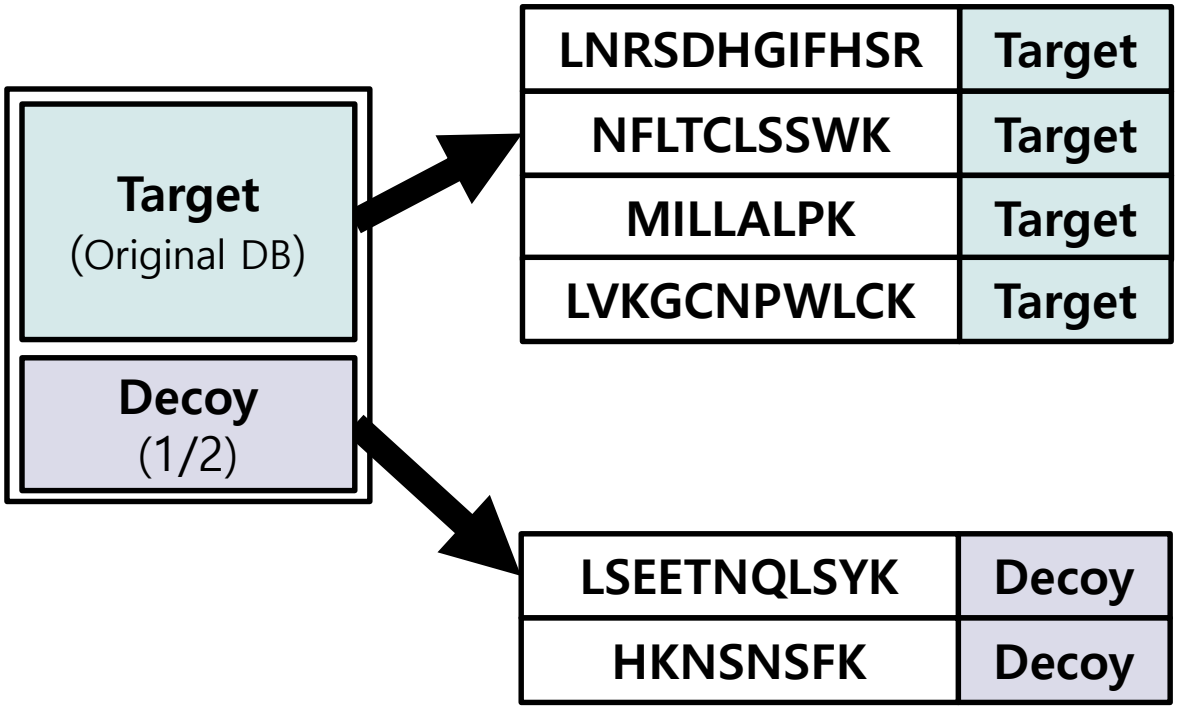




Target ?	Decoy ?
66%	33%

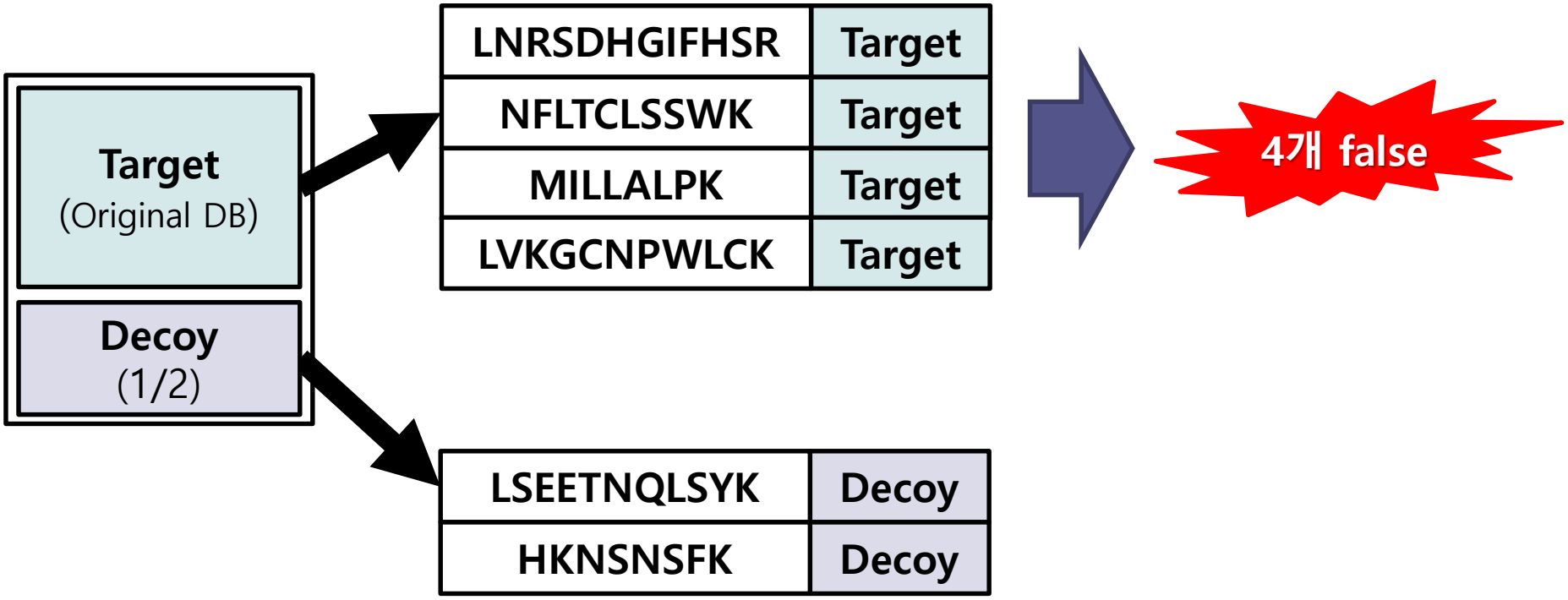
METHOD

- **FDR estimation using small decoy database**



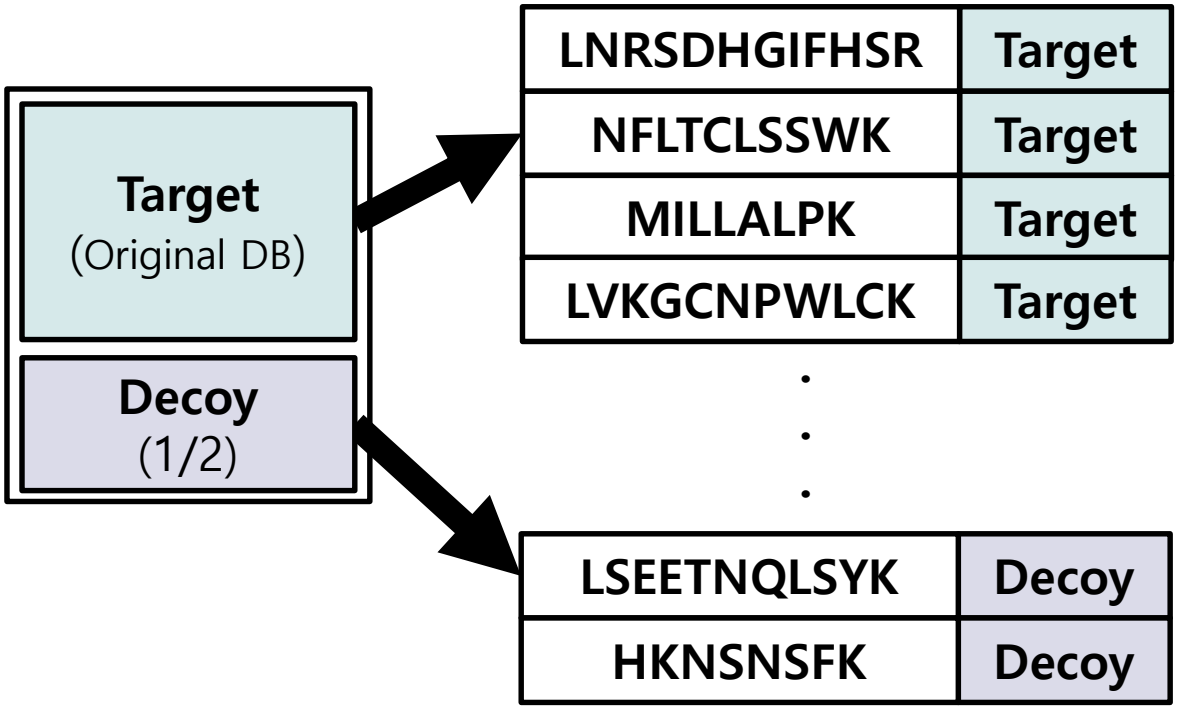
METHOD

- **FDR estimation using small decoy database**



METHOD

- **FDR estimation using small decoy database**



$$\frac{\#Decoy \times 2}{\#Target \times 1} = FDR$$

METHOD

- **Target-small decoy search**
  - How to calculate the probability of target and decoy false positives

METHOD

- **Target-small decoy search**
  - How to calculate the probability of target and decoy false positives

➔ Shifted MS/MS data

METHOD

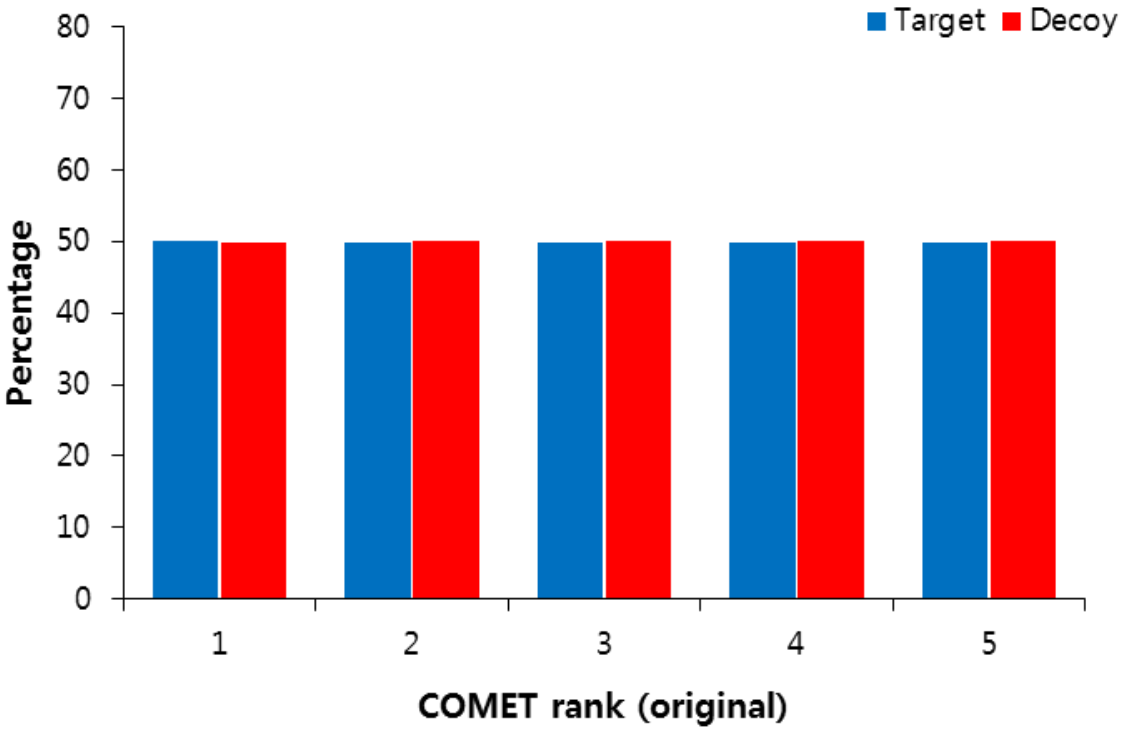
- **Target-small decoy search**
  - How to calculate the probability of target and decoy false positives

➔ Shifted MS/MS data

  - Precursor masses of MS/MS data are shifted by 10 Da

METHOD

- **Target-decoy search strategy**



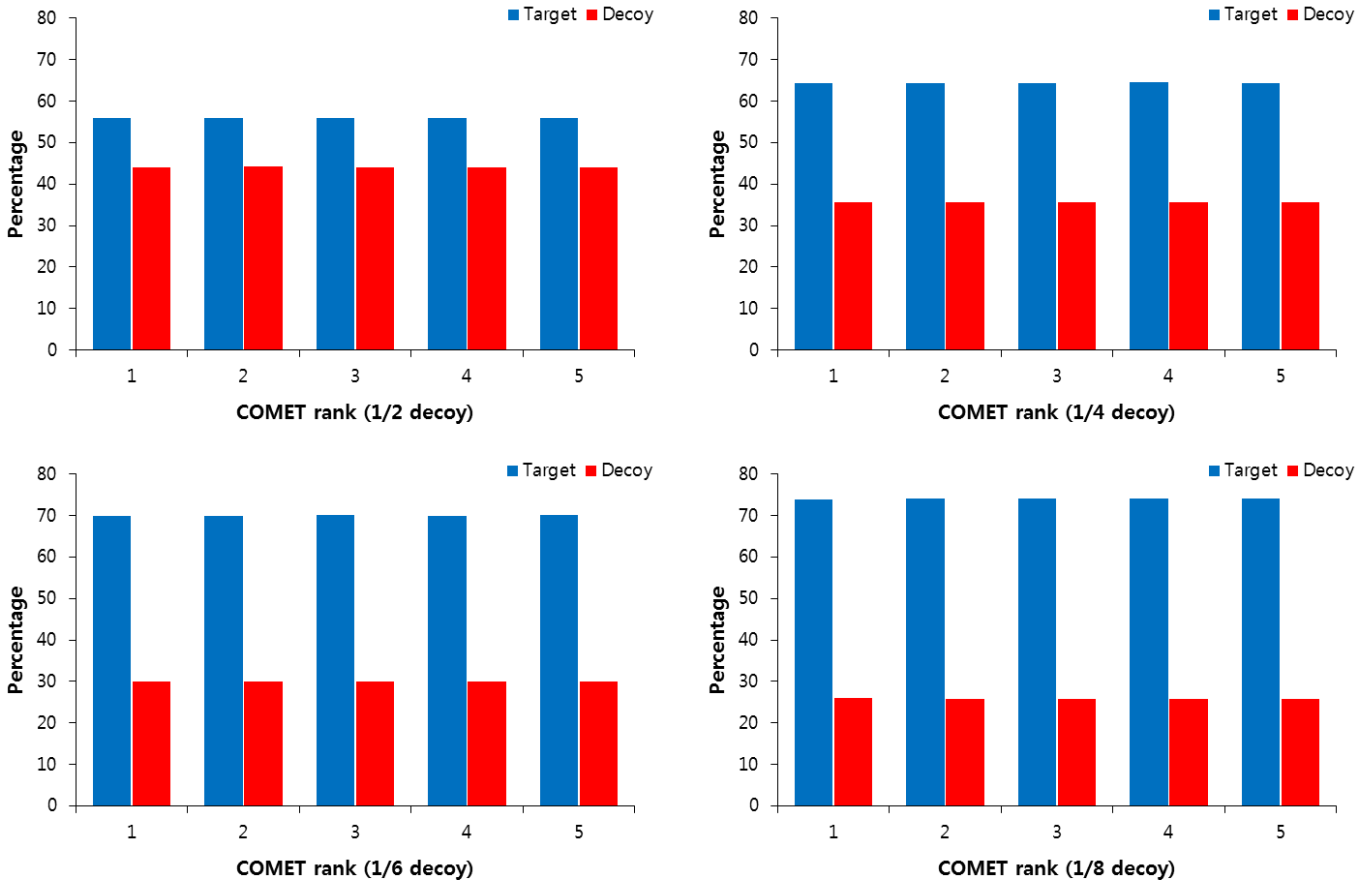
Comet

precursor mass tolerance = 10 ppm
fragment tolerance = 0.02 Da
number of tryptic termini (NTT) = 2
maximum missed cleavage = 2
fixed modification of carbamidomethyl on Cys

HEK293 cell line data

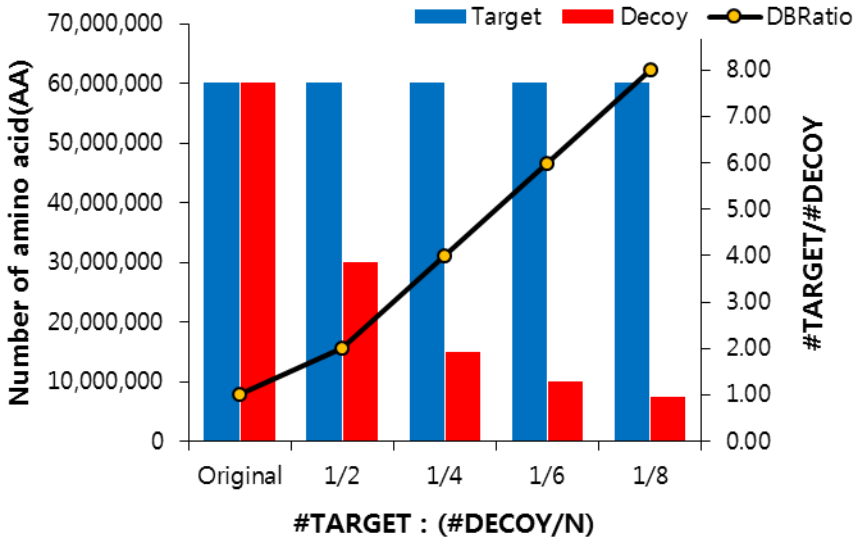
METHOD

- **Target-small decoy search strategy**

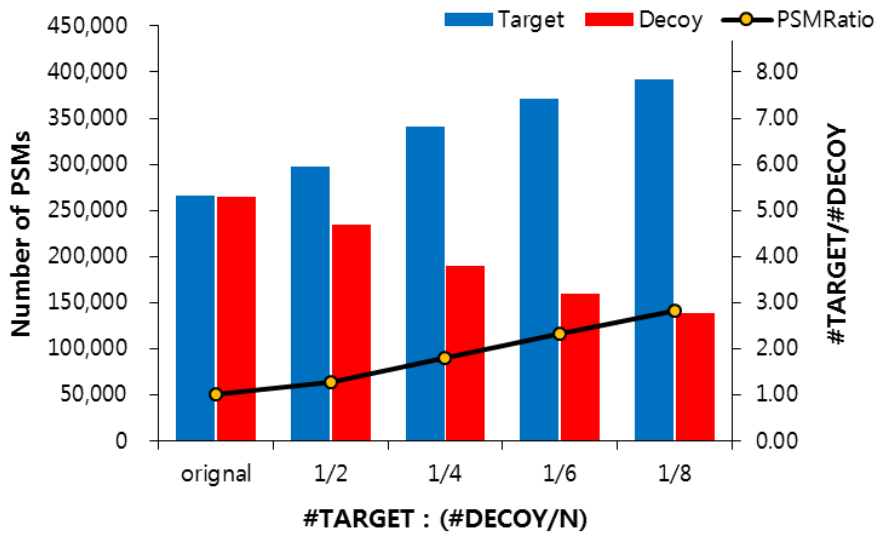


METHOD

- There is a difference
  - The ratio of the size of a target and a decoy database
  - The ratio of the number of target and decoy PSMs



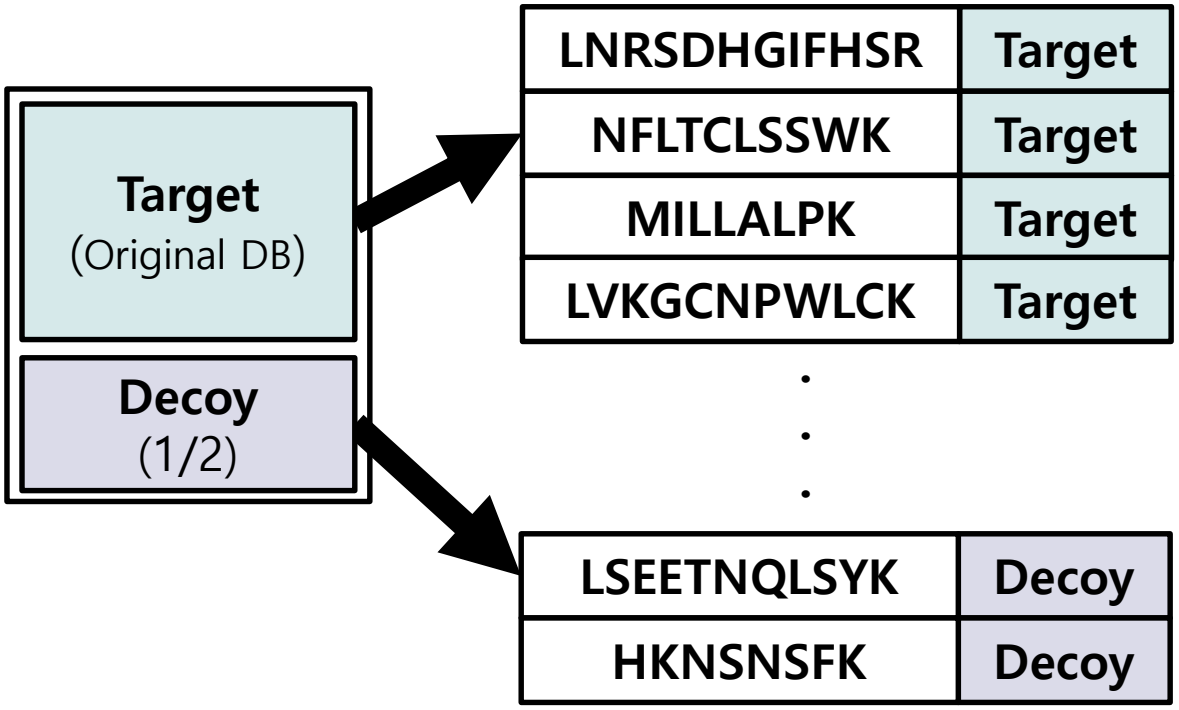
The size of database



The number of PSMs

METHOD

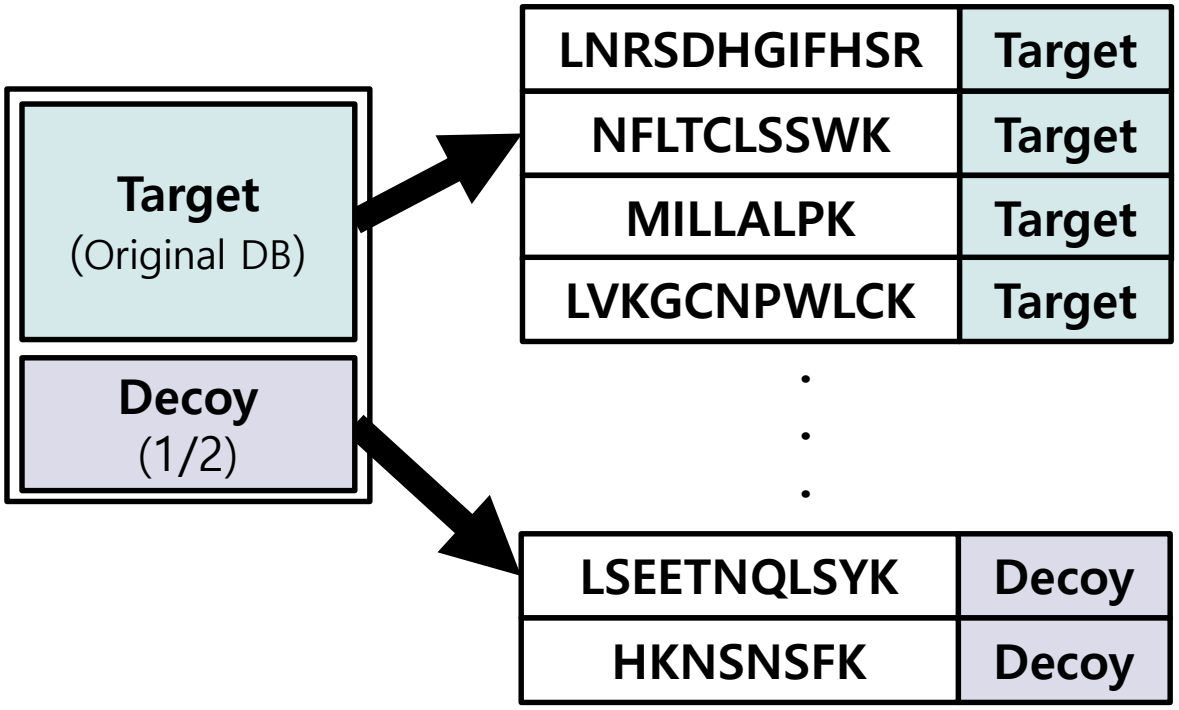
- **FDR estimation using small decoy database**



$$\frac{\#Decoy \times 1}{\#Target \times 2} = FDR$$

METHOD

- **FDR estimation using small decoy database**



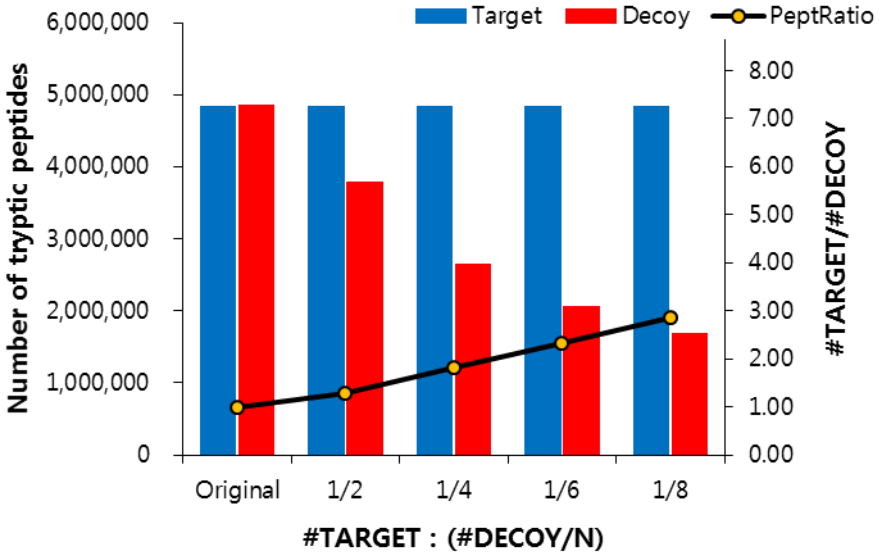
$$\frac{\#Decoy \times 1}{\#Target \times 2} = FDR$$

METHOD

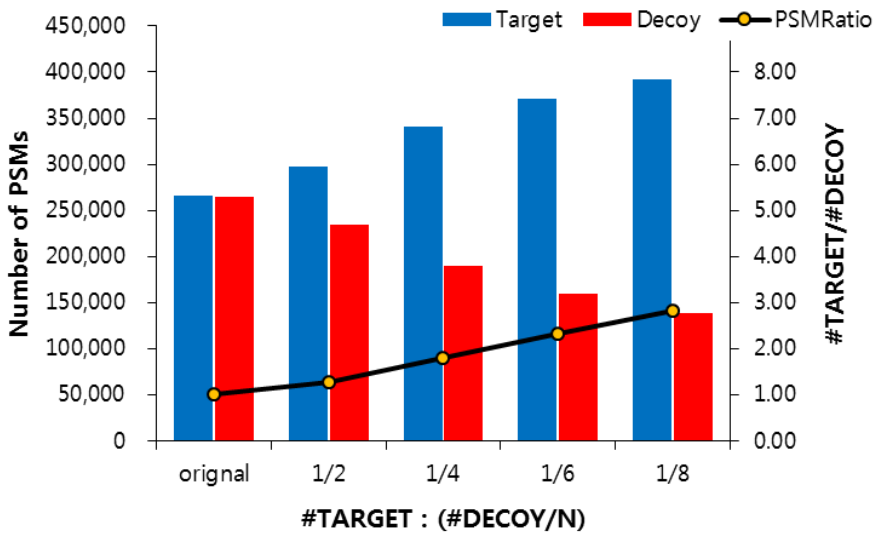
- **FDR estimation using small decoy database**
  - Candidate peptides within tolerance of precursor masses instead of proteins are used for peptide identification
  - Because of this reason, we expected that the ratio of target and decoy peptides in databases are almost same as the ratio of random hits

METHOD

- Almost the same
  - The ratio of the number of a target and a decoy peptides
  - The ratio of the number of target and decoy PSMs



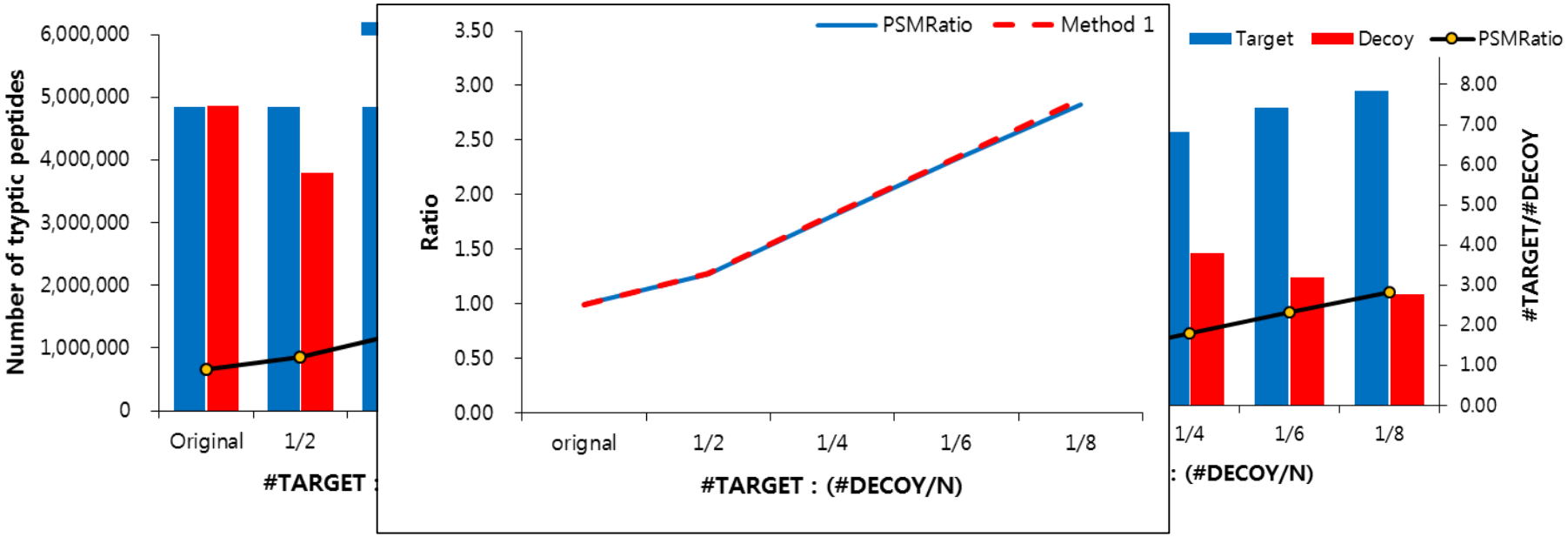
The number of peptides



The number of PSMs

METHOD

- Almost the same
  - The ratio of the number of a target and a decoy peptides
  - The ratio of the number of target and decoy PSMs



The number of peptides

The number of PSMs

METHOD

- **How to estimate the FDR**

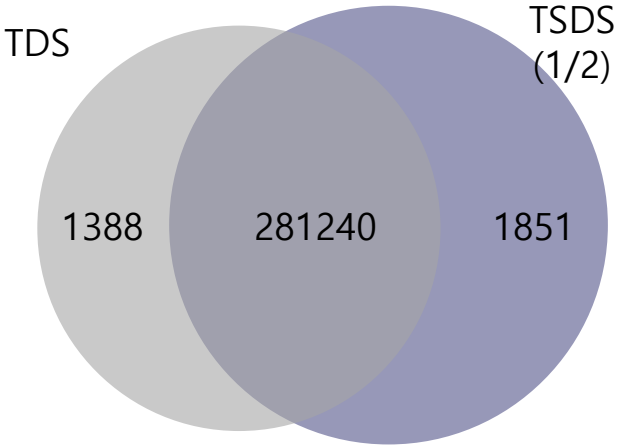
$$FDR_{TDS} = \frac{\#Decoy}{\#Target} \qquad FDR_{TSDS} = \frac{\#Decoy}{\#Target} \times FPratio$$

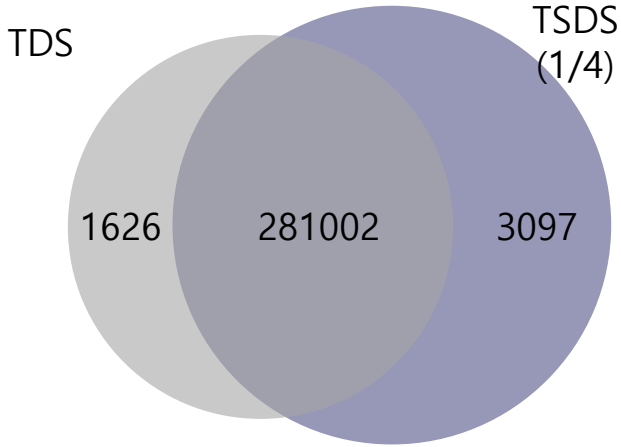
  - *#Target* is the number of target PSMs
  - *#Decoy* is the number of decoy PSMs
  - *FPratio* is the ratio of target and decoy peptides

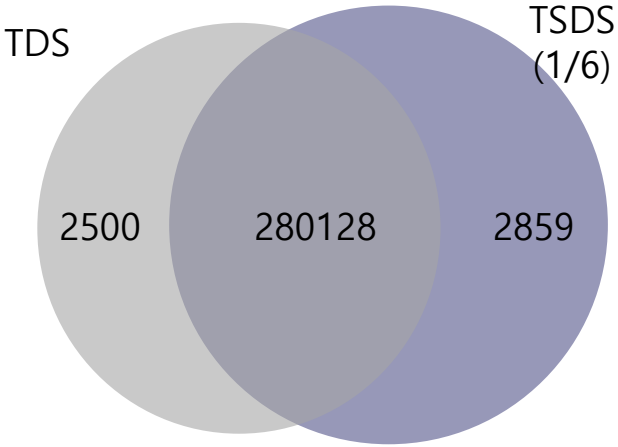
- **1% FDR**
  - Target-decoy search strategy: $FDR_{TDS} = 0.01$
  - Target-small decoy search strategy: $FDR_{TDS} = 0.01/FPratio$

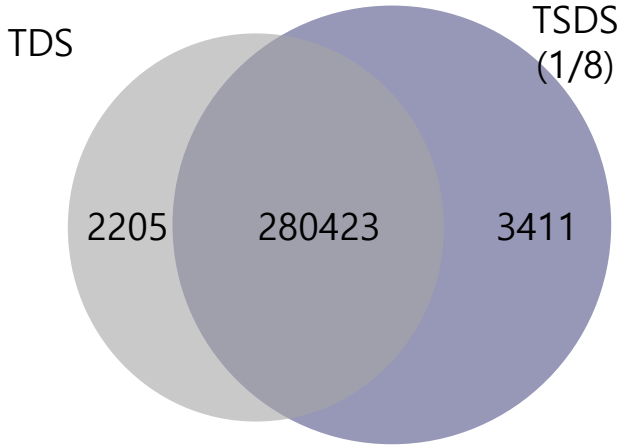
RESULT

- **Comparison PSMs**



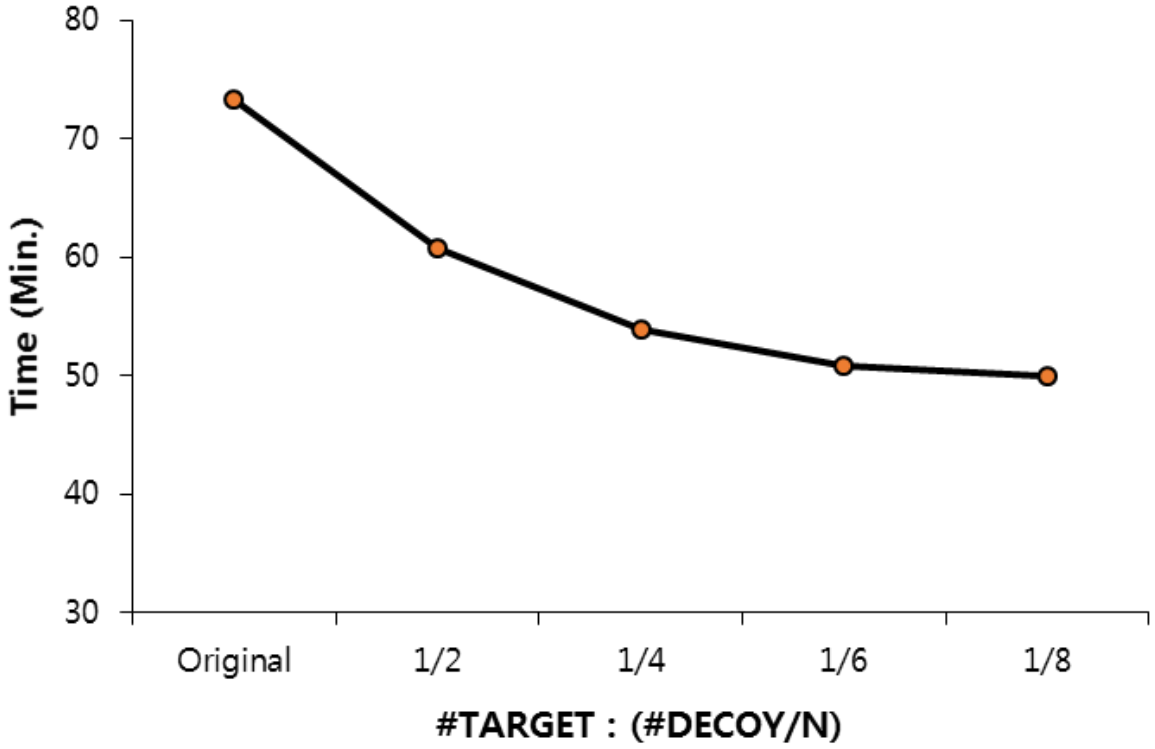






RESULT

- **Comparison time**



CONCLUSION

- **Target-small decoy search strategy**
  - Efficient as target-decoy search strategy
  - Easily estimate the FDR
  - Reduce the database search time

Q & A