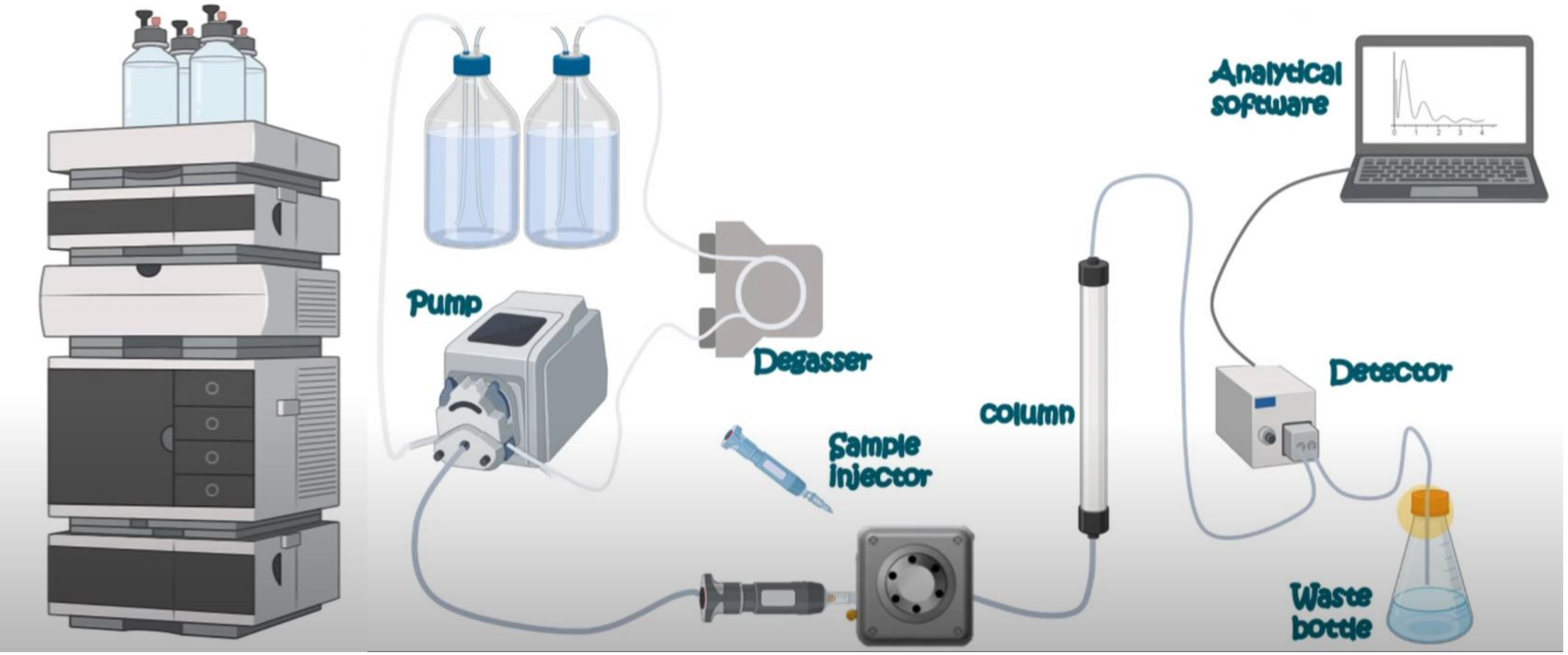
FineFDR:
Fine-grained Taxonomy-specific False Discovery Rates Control in Metaproteomics

Shengze Wang [1], Shichao Feng[1], Chongle Pan[2], Xuan Guo[1]

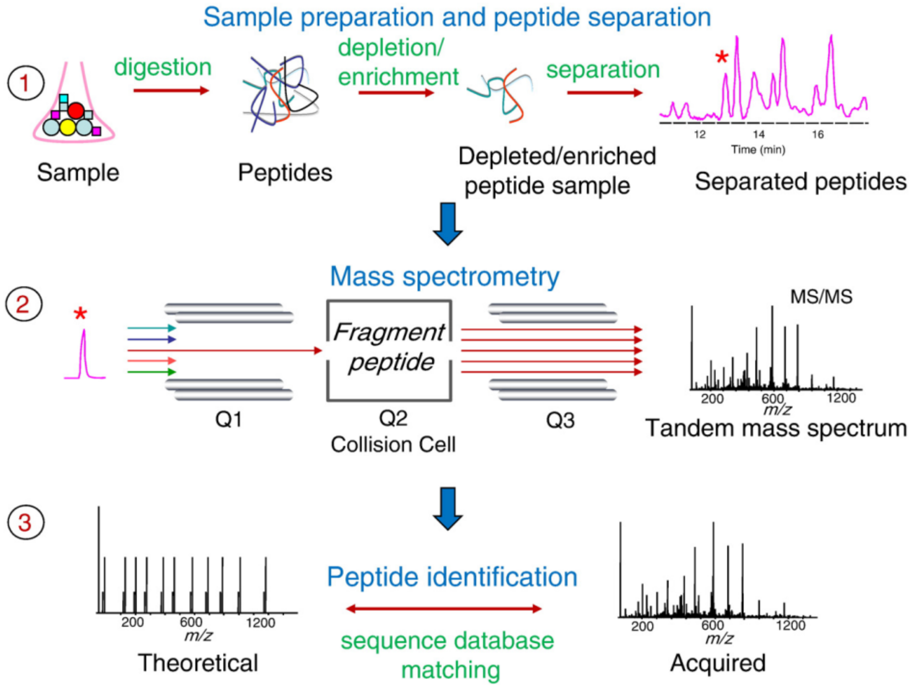
[1] Department of Computer Science and Engineering, University of North Texas
[2] School of Computer Science & Department of Microbiology and Plant Biology, University of Oklahoma

Background

1 **Current discovery metaproteomics studies** are generally based on high-throughput tandem mass spectrometry (MS/MS) coupled with liquid chromatography (LC). (<u>LC-MS/MS</u>)



Source: A. biology With arpan, "HPLC | High Performance Liquid Chromatography | Application of HPLC," 16-Sep-2020. [Online]. Available: https://www.youtube.com/watch?v=Vr5t-cgHHG4. [Accessed: 23-Nov-2022].

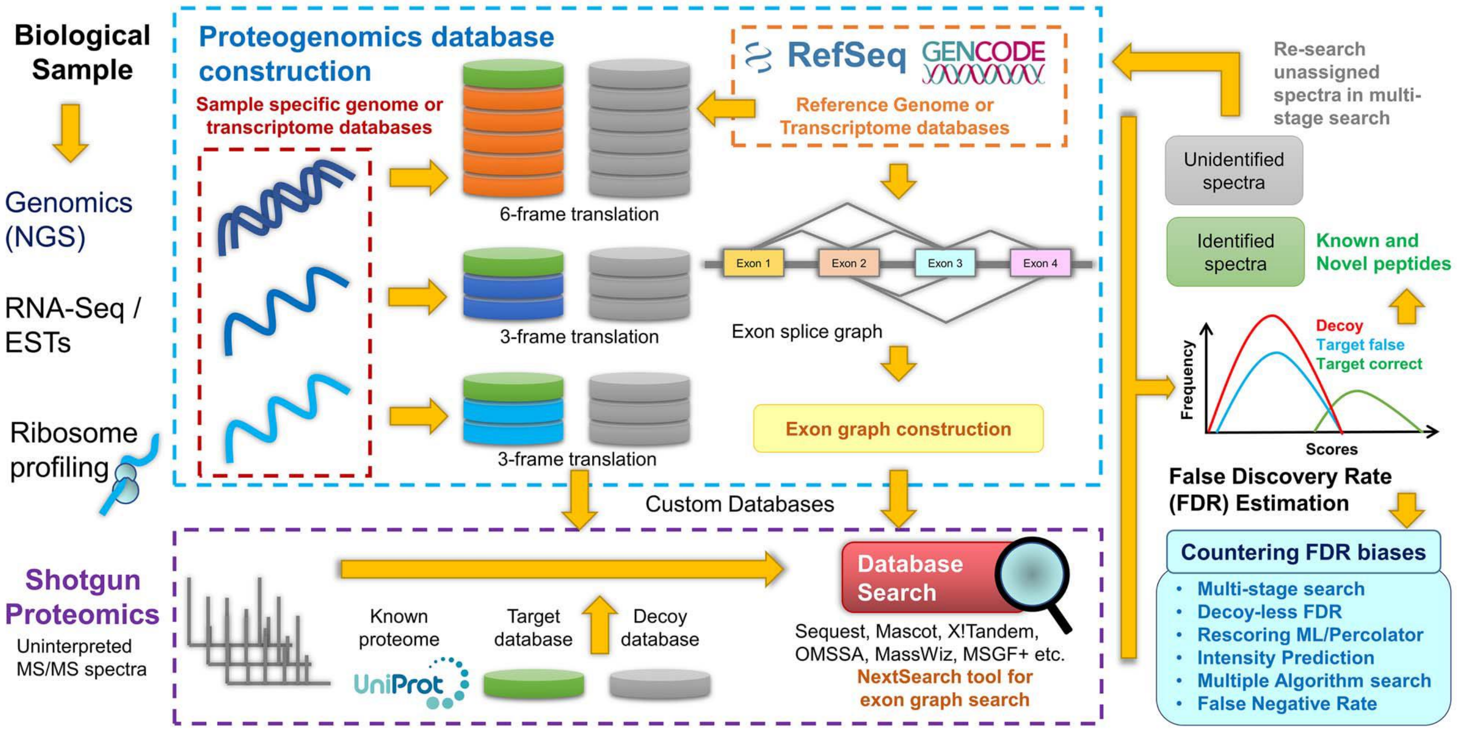


Source: A. I. Nesvizhskii, "A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics," Journal of Proteomics

Background

2

Identifying peptides and proteins from microbiota
involves a procedure of searching mass spectra against a pre-defined protein sequence database.



Source: S. Aggarwal, A. Raj, D. Kumar, D. Dash, and A. K. Yadav, “False discovery rate: the Achilles’ heel of proteogenomics,” Briefings in Bioinformatics, vol. 23, no. 5. Oxford University Press (OUP), May 09, 2022. DOI: 10.1093/bib/bbac163.

Background

3

A major post-analysis step

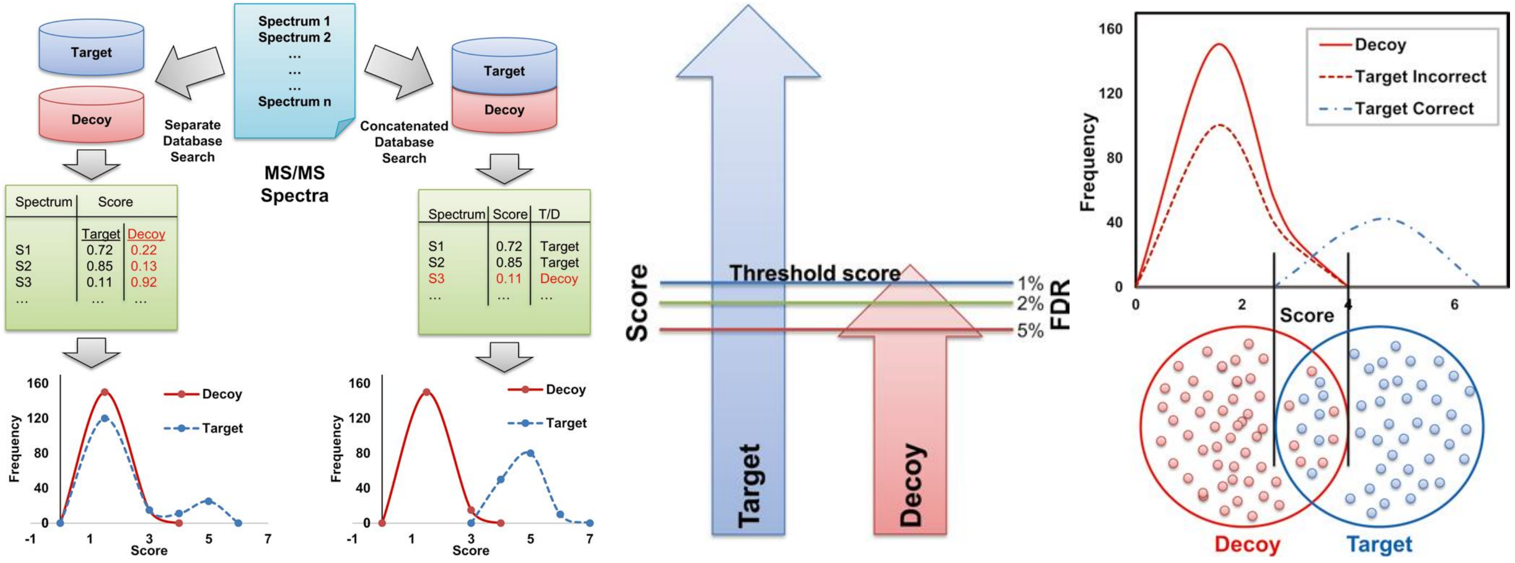
is controlling the **false discovery rate**, i.e.,

FDR, the ratio of false positives to the total number of annotations.

4

The current gold standard for FDR estimation

is the target-decoy search strategy using p-value or E-value.



Spectrum	Score	
	Target	Decoy
S1	0.72	0.22
S2	0.85	0.13
S3	0.11	0.92
...	...	...

Spectrum	Score	T/D
S1	0.72	Target
S2	0.85	Target
S3	0.11	Decoy
...	...	...

$$FDR = \frac{\#\,Decoys}{\#\,Targets}$$

Source: S. Aggarwal and A. K. Yadav, “False Discovery Rate Estimation in Proteomics,” Methods in Molecular Biology. Springer New York, pp. 119–128, 2016. doi: 10.1007/978-1-4939-3106-4_7.

Motivation

Reliability of identifications

a. Single-identification level

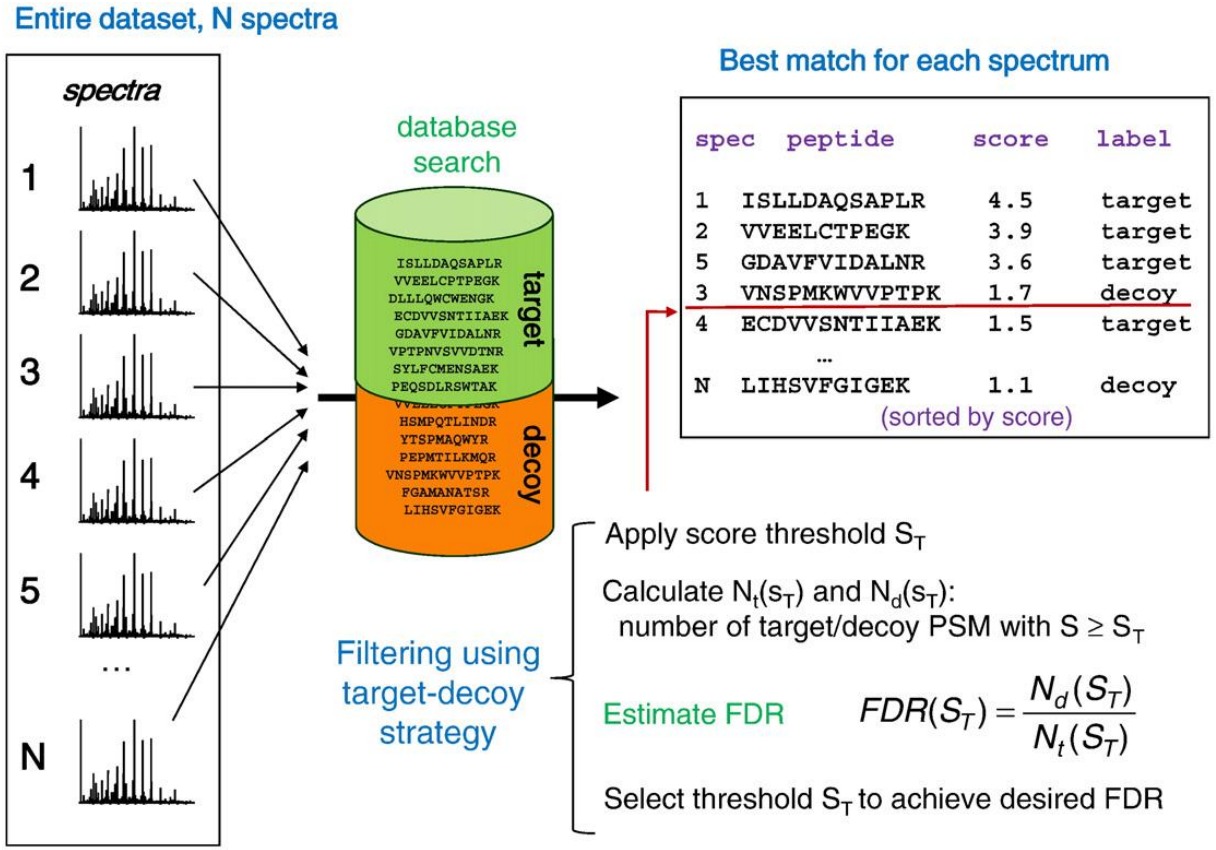
p-value or E-value

b. Multiple-identification level

The proportion of incorrect identifications for a group of identifications.

The problem of FDR estimation in multiple hypothesis tests:

Treat all the peptides and proteins equally and overlook that they could have varied probabilities of being identified.



Best match for each spectrum

spec	peptide	score	label
1	ISLLDAQSAPLR	4.5	target
2	VVEELCTPEGK	3.9	target
5	GDAVFVIDALNR	3.6	target
3	VNSPMKWVVPTPK	1.7	decoy
4	ECDVVSNTIIAEK	1.5	target
	…		
N	LIHSVFGIGEK	1.1	decoy

(sorted by score)

Apply score threshold S_T

Calculate $N_t(s_T)$ and $N_d(s_T)$: number of target/decoy PSM with $S \geq S_T$

Estimate FDR $$FDR(S_T) = \frac{N_d(S_T)}{N_t(S_T)}$$

Select threshold S_T to achieve desired FDR

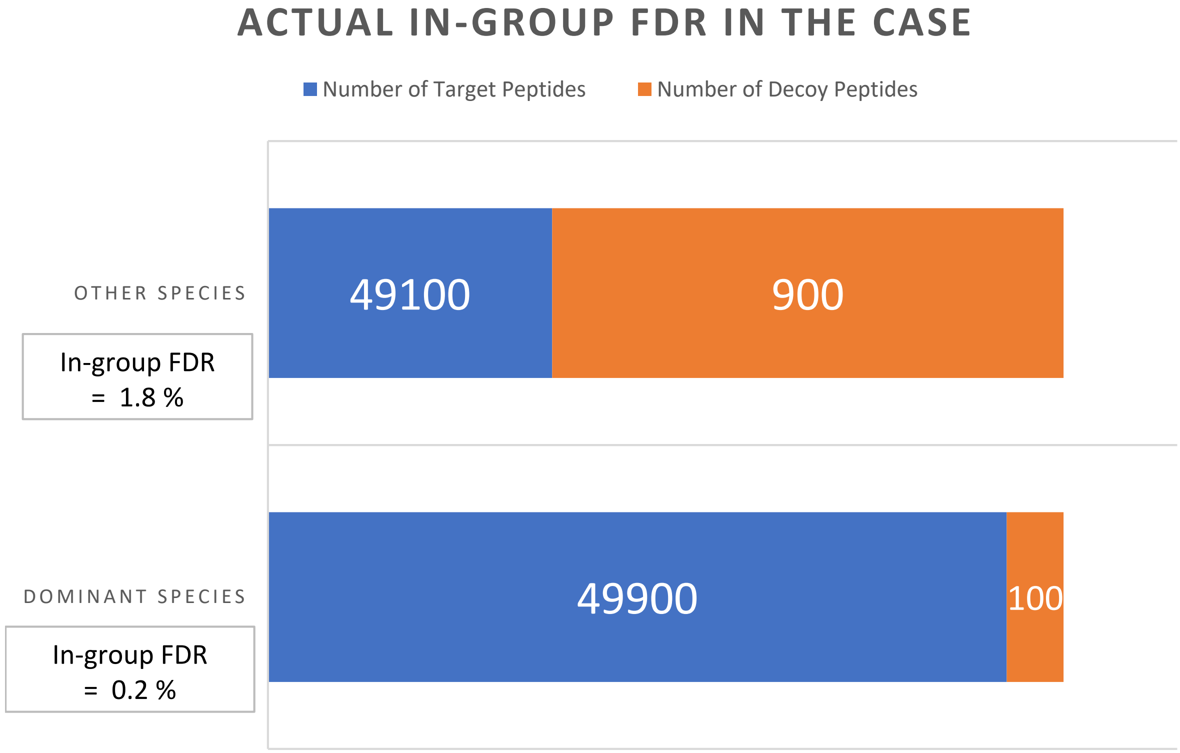
Source: A. I. Nesvizhskii, “A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics,” Journal of Proteomics, vol. 73, no. 11. Elsevier BV, pp. 2092–2123, Oct. 2010.

Motivation

The problem of FDR estimation in multiple hypothesis tests:

Treat all the peptides and proteins equally and overlook that they could have varied probabilities of being identified.

In an extreme case, If we have
50,000 identified peptides from a dominant species,
50,000 identified peptides from other species;
FDR level is set to be 1%,
so expected false-positive identification = 100,000 x 0.01 = 1,000;
Varied probabilities of being identified:
10% of false-positive were from the dominant species, and the left was from the other species.



In-group FDR = 1.8 %

In-group FDR = 0.2 %

Main idea

FineFDR controls the FDR separately for PSMs/peptides/proteins from the different taxonomic units.

Assumption

Peptides and proteins are not equally likely to be measured by LC-MS/MS and identified by search engines due to the varied abundance of microorganisms.

Method: Target-decoy FDR Control

The basic target-decoy strategy augments the "target" protein database with a set of “decoy” protein sequences.

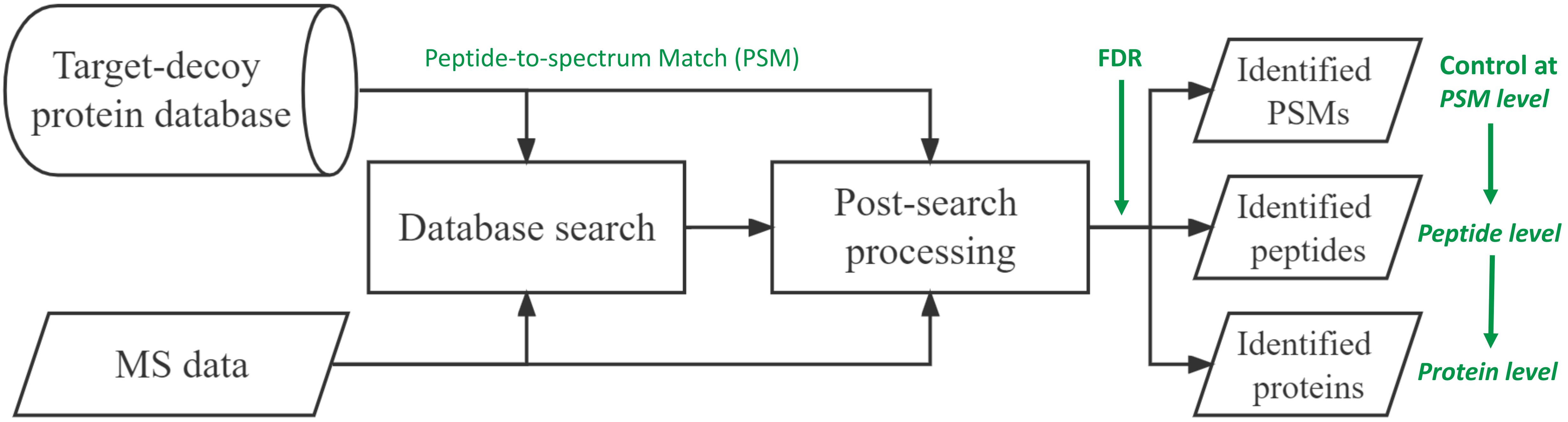


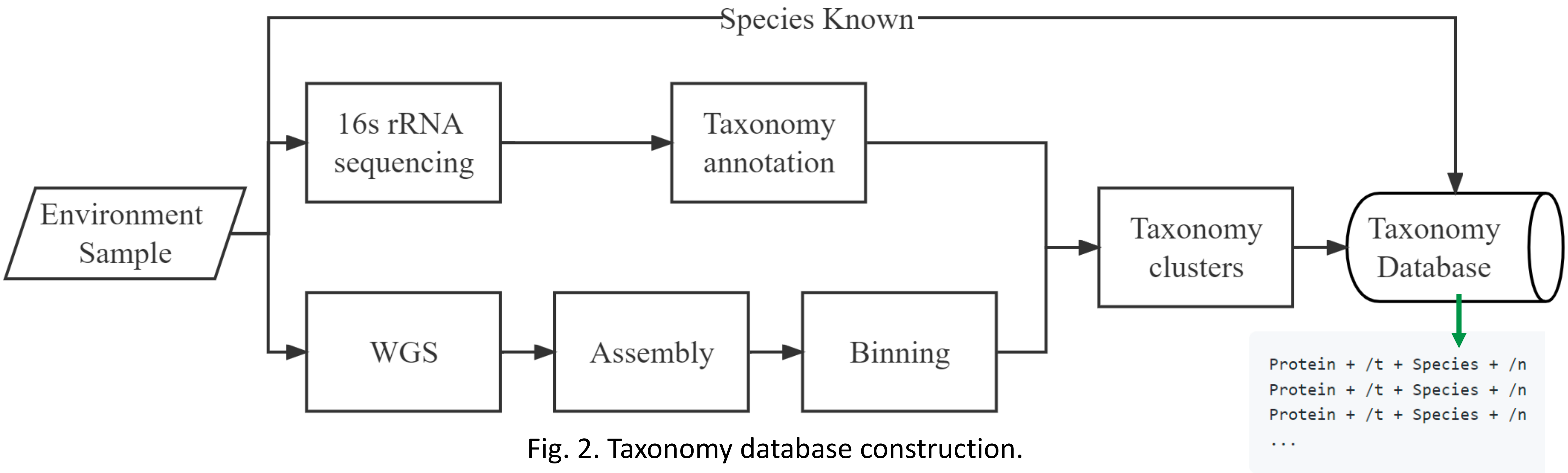
Fig. 1. The basic workflow of the target-decoy search strategy.

Method: Taxonomy Database Construction

Operational Taxonomic Unit (OTU)

- Groups of closely related microorganisms at the genome level
- Basic unit to group PSMs or peptides in FineFDR

Peptide-to-Spectrum Matches (PSM)



```
Protein + /t + Species + /n
Protein + /t + Species + /n
Protein + /t + Species + /n
...
```

Fig. 2. Taxonomy database construction.

Method: Taxonomy-specific FDR assessment (PSM)

$$In - group\ FDR_{OTU\ index\ i} = \frac{\#\ Decoys \subset OTU\ Cluster_i}{\#\ Targets \subset OTU\ Cluster_i}\ ,\quad Gobal\ FDR_i = \frac{\#\ Decoys}{\#\ Targets}$$

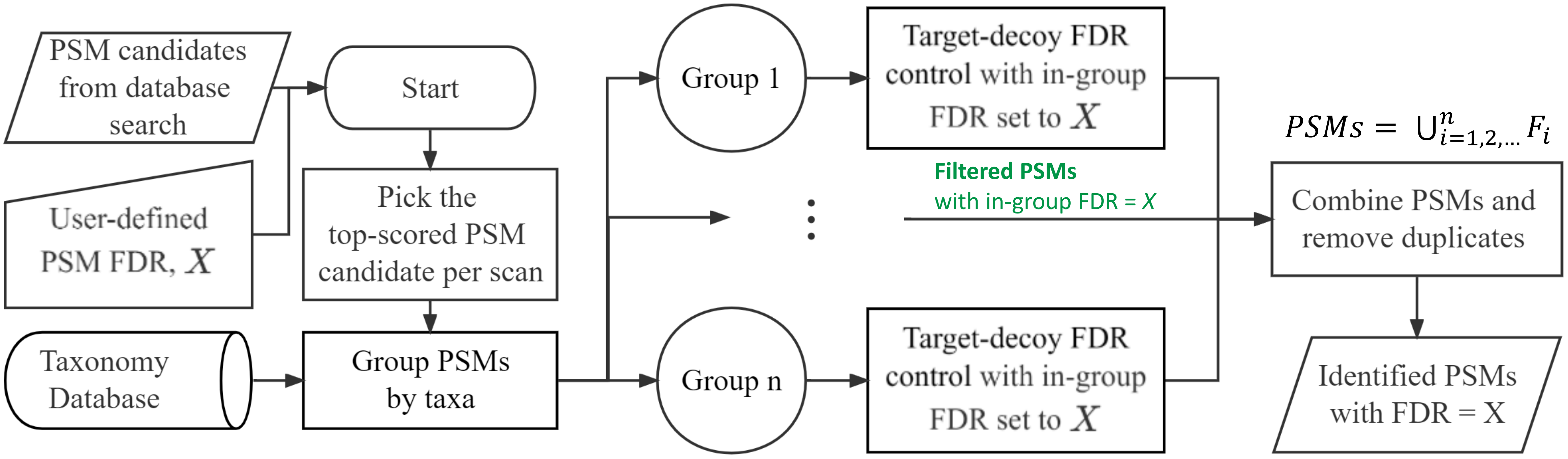
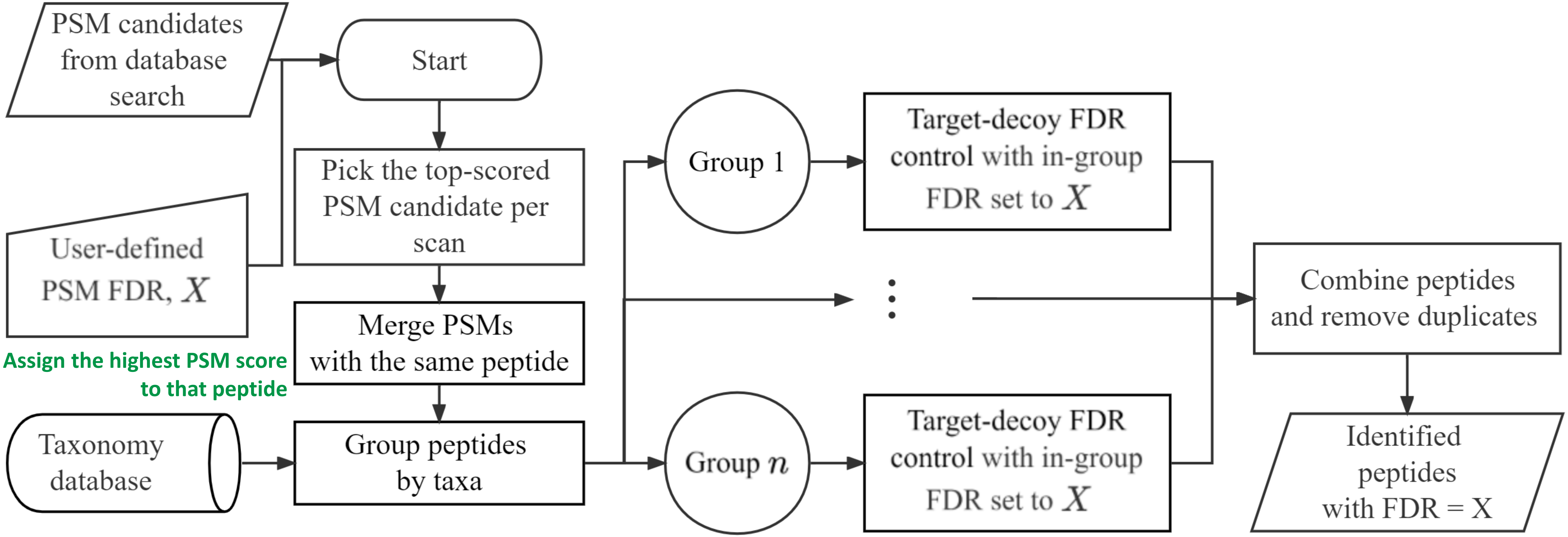

PSM candidates from database search
User-defined PSM FDR, X
Taxonomy Database
Start
Pick the top-scored PSM candidate per scan
Group PSMs by taxa
Group 1
Group n
Target-decoy FDR control with in-group FDR set to X
Target-decoy FDR control with in-group FDR set to X
Filtered PSMs
with in-group FDR = X
$PSMs = \cup_{i=1,2,\ldots}^{n} F_i$
Combine PSMs and remove duplicates
Identified PSMs with FDR = X

Fig. 3. The framework of taxonomy-specific FDR control at the PSM level.

Method: Taxonomy-specific FDR assessment (Peptide)


PSM candidates from database search
Start
User-defined PSM FDR, X
Pick the top-scored PSM candidate per scan
Merge PSMs with the same peptide
Taxonomy database
Group peptides by taxa
Group 1
Group n
Target-decoy FDR control with in-group FDR set to X
Target-decoy FDR control with in-group FDR set to X
Combine peptides and remove duplicates
Identified peptides with FDR = X

Assign the highest PSM score to that peptide

Fig. 4. The framework of taxonomy-specific FDR control at the peptide level.

Method: Taxonomy-specific FDR assessment (Protein)

Because transferring PSM scores to protein scores is not trivial, FineFDR adjusts in-group peptide and protein FDRs dynamically *Until* the global protein FDR is well controlled.

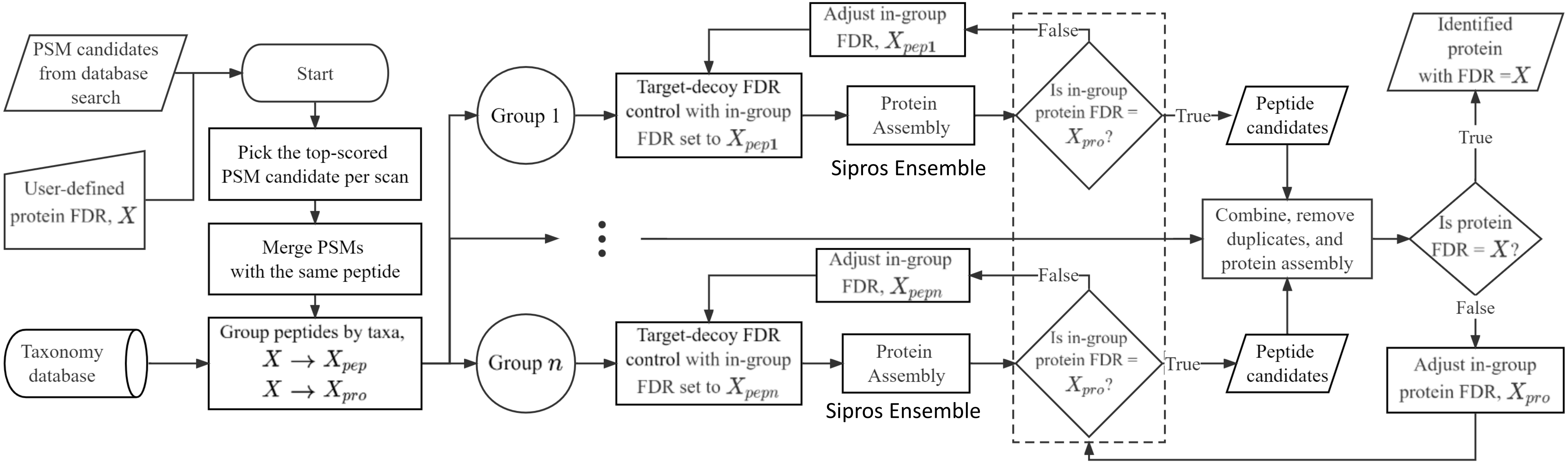

PSM candidates from database search
User-defined protein FDR, X
Taxonomy database
Start
Pick the top-scored PSM candidate per scan
Merge PSMs with the same peptide
Group peptides by taxa, $X \rightarrow X_{pep}$ $X \rightarrow X_{pro}$
Group 1
Group n
Target-decoy FDR control with in-group FDR set to X_{pep1}
Target-decoy FDR control with in-group FDR set to X_{pepn}
Adjust in-group FDR, X_{pep1}
Adjust in-group FDR, X_{pepn}
Protein Assembly
Sipros Ensemble
Protein Assembly
Sipros Ensemble
False
Is in-group protein FDR = X_{pro}?
True
False
Is in-group protein FDR = X_{pro}?
True
Peptide candidates
Peptide candidates
Combine, remove duplicates, and protein assembly
Is protein FDR = X?
True
False
Identified protein with FDR =X
Adjust in-group protein FDR, X_{pro}

Fig. 5. The framework of taxonomy-specific FDR control at the protein level.

Experiment Design

Search Tool

- Comet (E-value score; Widely-used method)

Filtering Tool

- Percolator (Percolator score; Widely-used method)
- TIDD (TIDD SVM Prob; Recent work)
- Tailor (Tailor score; Recent work)

Data sets

- Simulated data set: Mock microbial ”U” (UNEVEN) type community data set with the cell number U1 (PXD006118)
- Real-world data set: Marine 1,2,3 (PXD007587); Soil 1,2,3 (PXD007587); Human Gut (PXD013386)
- Simulated ground-truth data set: GT

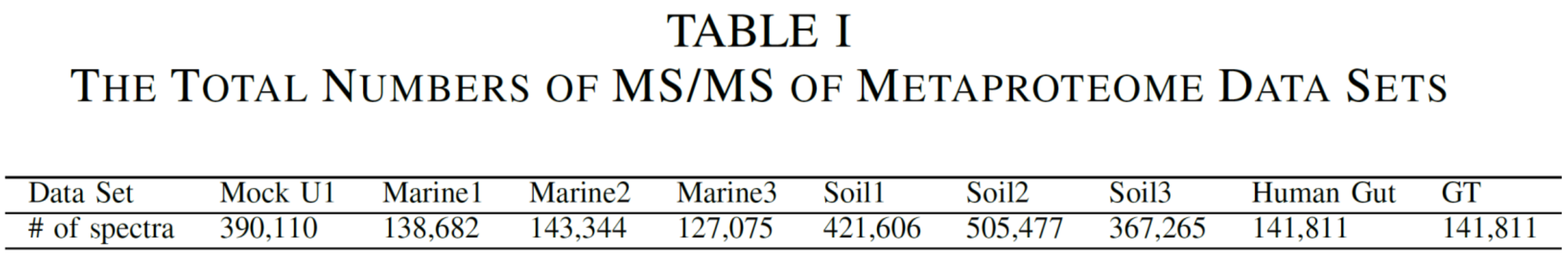


TABLE I
THE TOTAL NUMBERS OF MS/MS OF METAPROTEOME DATA SETS

Data Set	Mock U1	Marine1	Marine2	Marine3	Soil1	Soil2	Soil3	Human Gut	GT
# of spectra	390,110	138,682	143,344	127,075	421,606	505,477	367,265	141,811	141,811

Identification Quality

$$Precision = \frac{\#True\ identifications}{\#Identifications}$$

Ground-truth data set simulation

Search the Human Gut's MS data against the database consisting of the protein mixture from the Human Gut and Marine protein databases.

"Truth" identifications

The PSMs/peptides/proteins from the Human Gut proteome.

Result

FineFDR achieved higher precision than the baseline methods.

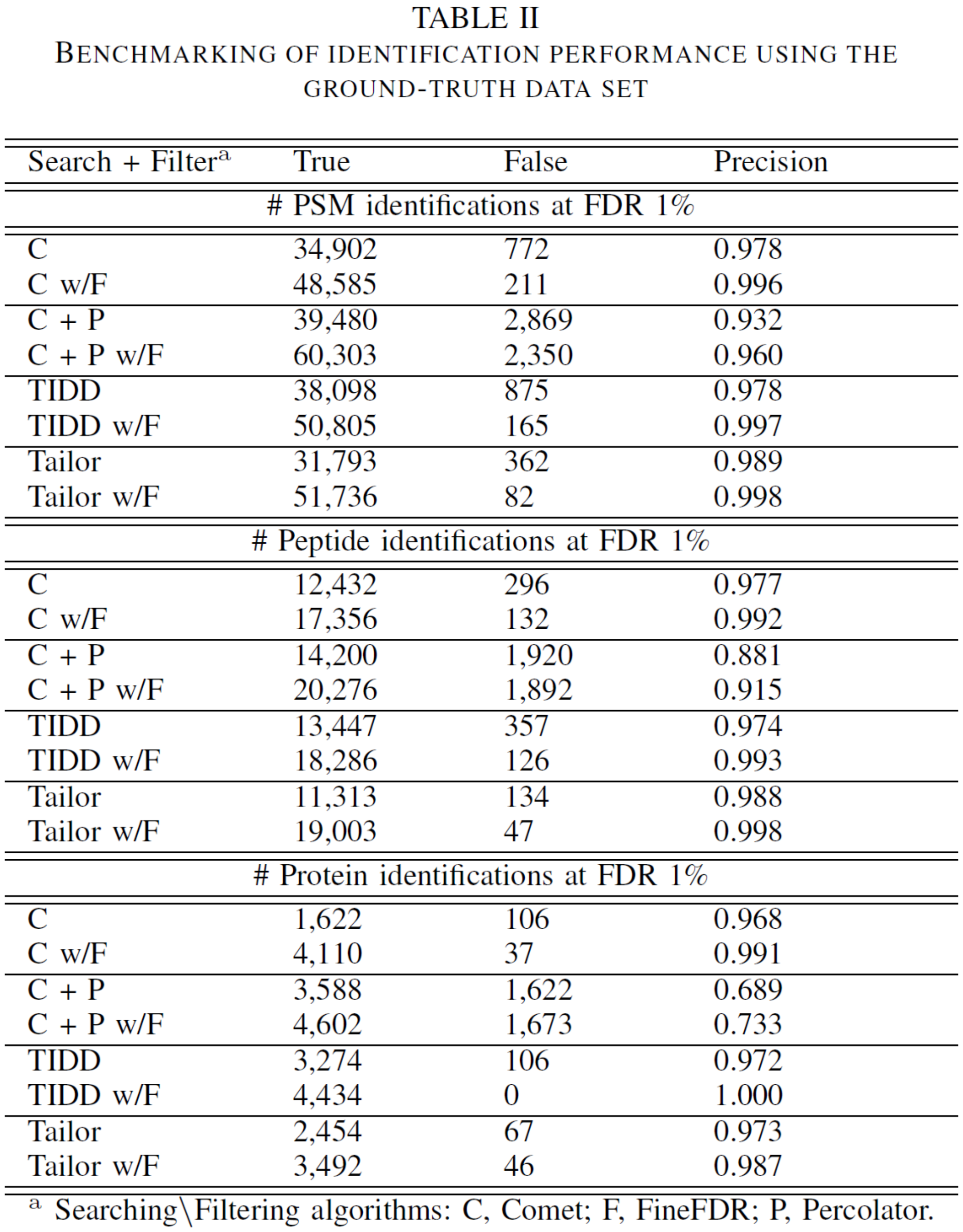


TABLE II
BENCHMARKING OF IDENTIFICATION PERFORMANCE USING THE GROUND-TRUTH DATA SET

Search + Filter[a]	True	False	Precision
# PSM identifications at FDR 1%			
C	34,902	772	0.978
C w/F	48,585	211	0.996
C + P	39,480	2,869	0.932
C + P w/F	60,303	2,350	0.960
TIDD	38,098	875	0.978
TIDD w/F	50,805	165	0.997
Tailor	31,793	362	0.989
Tailor w/F	51,736	82	0.998
# Peptide identifications at FDR 1%			
C	12,432	296	0.977
C w/F	17,356	132	0.992
C + P	14,200	1,920	0.881
C + P w/F	20,276	1,892	0.915
TIDD	13,447	357	0.974
TIDD w/F	18,286	126	0.993
Tailor	11,313	134	0.988
Tailor w/F	19,003	47	0.998
# Protein identifications at FDR 1%			
C	1,622	106	0.968
C w/F	4,110	37	0.991
C + P	3,588	1,622	0.689
C + P w/F	4,602	1,673	0.733
TIDD	3,274	106	0.972
TIDD w/F	4,434	0	1.000
Tailor	2,454	67	0.973
Tailor w/F	3,492	46	0.987

[a] Searching\Filtering algorithms: C, Comet; F, FineFDR; P, Percolator.

Identification Rate

Identification Rate

Number of PSMs, peptides, and proteins filtered at 1% FDR

Result

For the methods adding FineFDR, they achieved more identifications than the baseline methods without FineFDR.

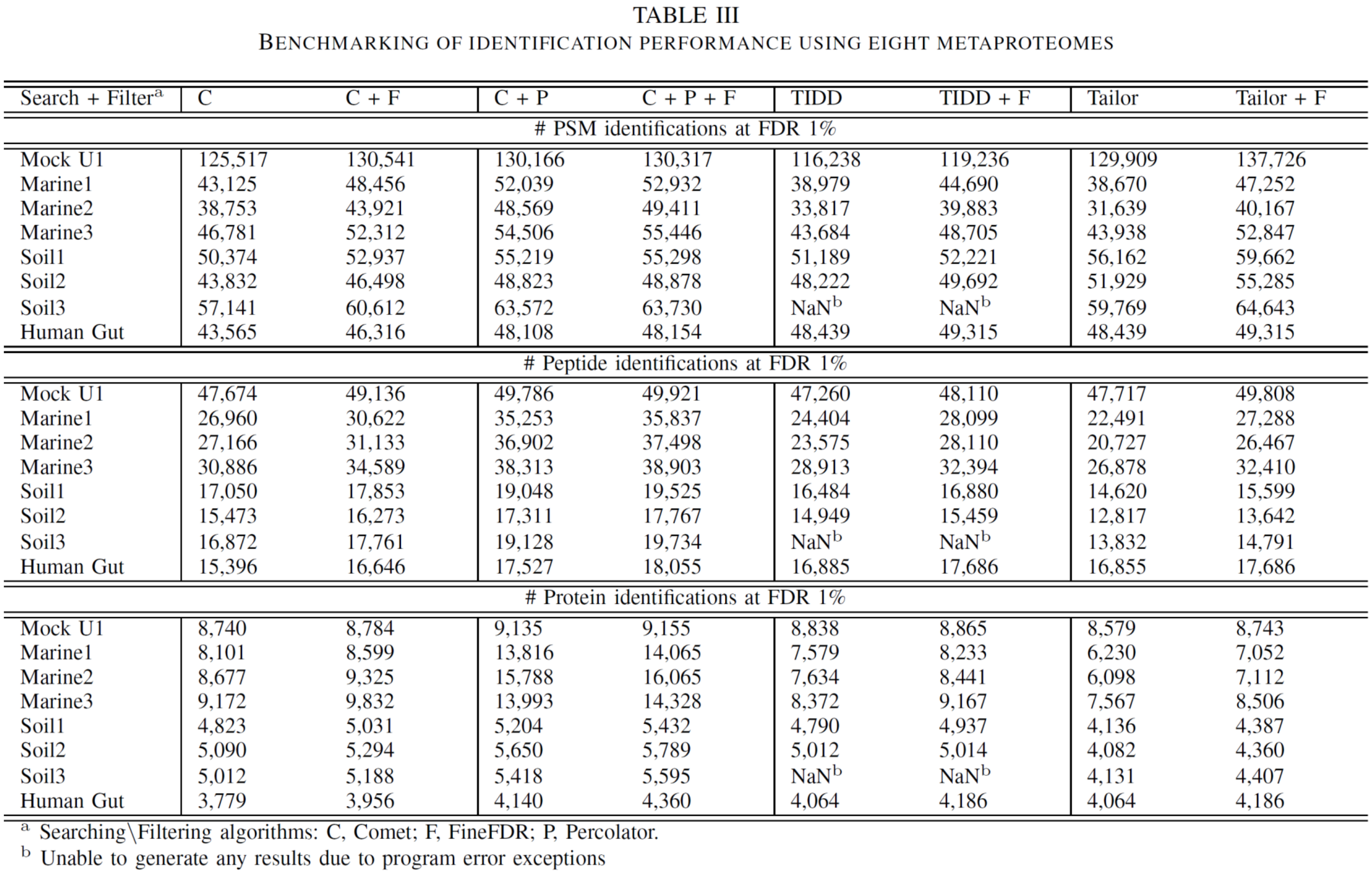


TABLE III

BENCHMARKING OF IDENTIFICATION PERFORMANCE USING EIGHT METAPROTEOMES

Search + Filter[a]	C	C + F	C + P	C + P + F	TIDD	TIDD + F	Tailor	Tailor + F
# PSM identifications at FDR 1%								
Mock U1	125,517	130,541	130,166	130,317	116,238	119,236	129,909	137,726
Marine1	43,125	48,456	52,039	52,932	38,979	44,690	38,670	47,252
Marine2	38,753	43,921	48,569	49,411	33,817	39,883	31,639	40,167
Marine3	46,781	52,312	54,506	55,446	43,684	48,705	43,938	52,847
Soil1	50,374	52,937	55,219	55,298	51,189	52,221	56,162	59,662
Soil2	43,832	46,498	48,823	48,878	48,222	49,692	51,929	55,285
Soil3	57,141	60,612	63,572	63,730	NaN[b]	NaN[b]	59,769	64,643
Human Gut	43,565	46,316	48,108	48,154	48,439	49,315	48,439	49,315
# Peptide identifications at FDR 1%								
Mock U1	47,674	49,136	49,786	49,921	47,260	48,110	47,717	49,808
Marine1	26,960	30,622	35,253	35,837	24,404	28,099	22,491	27,288
Marine2	27,166	31,133	36,902	37,498	23,575	28,110	20,727	26,467
Marine3	30,886	34,589	38,313	38,903	28,913	32,394	26,878	32,410
Soil1	17,050	17,853	19,048	19,525	16,484	16,880	14,620	15,599
Soil2	15,473	16,273	17,311	17,767	14,949	15,459	12,817	13,642
Soil3	16,872	17,761	19,128	19,734	NaN[b]	NaN[b]	13,832	14,791
Human Gut	15,396	16,646	17,527	18,055	16,885	17,686	16,855	17,686
# Protein identifications at FDR 1%								
Mock U1	8,740	8,784	9,135	9,155	8,838	8,865	8,579	8,743
Marine1	8,101	8,599	13,816	14,065	7,579	8,233	6,230	7,052
Marine2	8,677	9,325	15,788	16,065	7,634	8,441	6,098	7,112
Marine3	9,172	9,832	13,993	14,328	8,372	9,167	7,567	8,506
Soil1	4,823	5,031	5,204	5,432	4,790	4,937	4,136	4,387
Soil2	5,090	5,294	5,650	5,789	5,012	5,014	4,082	4,360
Soil3	5,012	5,188	5,418	5,595	NaN[b]	NaN[b]	4,131	4,407
Human Gut	3,779	3,956	4,140	4,360	4,064	4,186	4,064	4,186

[a] Searching\Filtering algorithms: C, Comet; F, FineFDR; P, Percolator.

[b] Unable to generate any results due to program error exceptions

Computational time

Test Platform

A regular desktop with an 8-Core 4.0 GHz CPU, 32GB 3200 MHz RAM, and NVMe 3.0 SSD.

FineFDR is implemented with Python 3.9.
On average, FineFDR requires 2 GB of memory to load data.

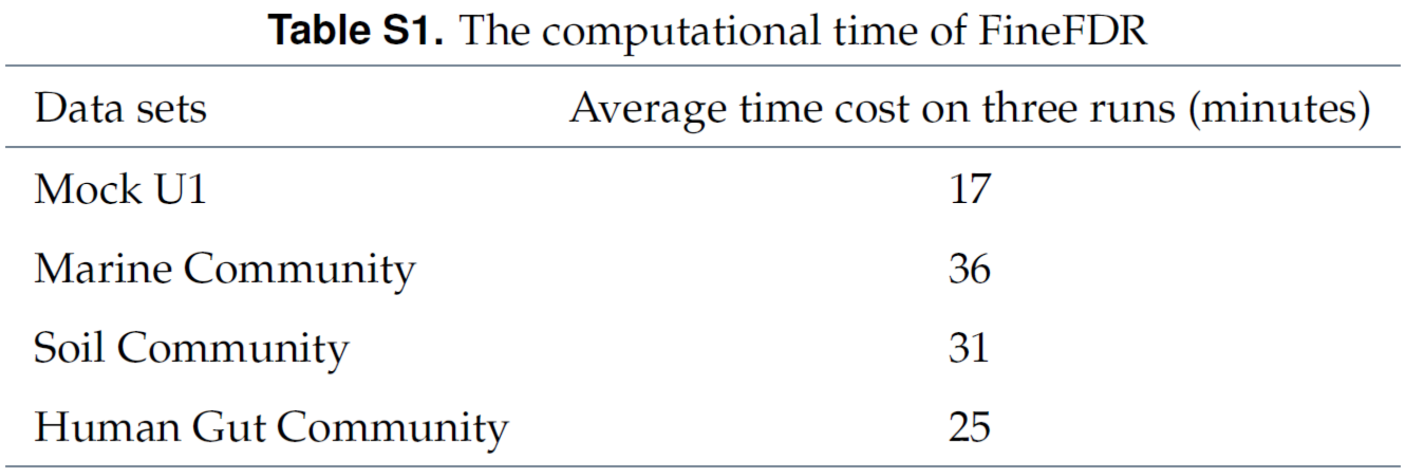


Table S1. The computational time of FineFDR

Data sets	Average time cost on three runs (minutes)
Mock U1	17
Marine Community	36
Soil Community	31
Human Gut Community	25

Discussion

1

The baseline method and its combination with FineFDR shared over 95% identical PSMs, peptides, and proteins in the results.
FineFDR made more method-specific discoveries than the baseline method.

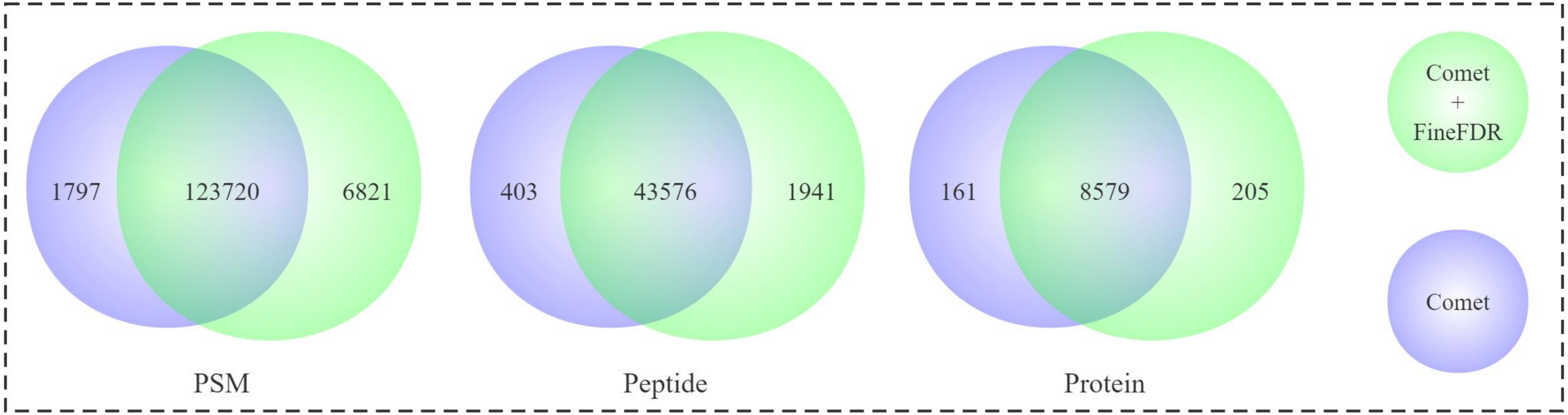


Fig. 6. The identified result overlap between the baseline method comet and its combination with FineFDR for the Mock U1

Discussion

2

FineFDR improved the identification rates across most species.

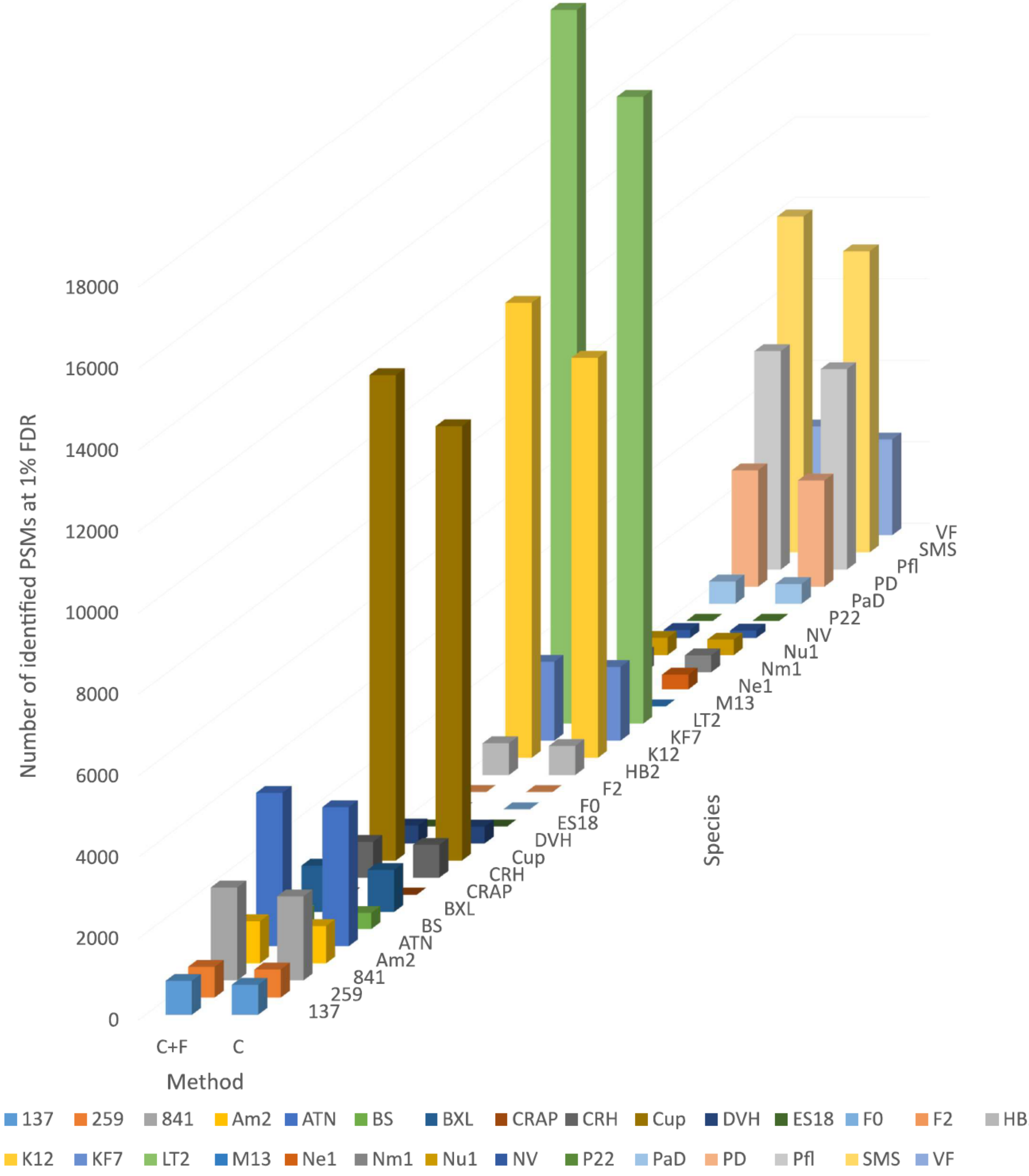


Fig. S1. PSM identification improvements by species for the Mock U1

Discussion

3

FineFDR shows the power to promote the percentage of target PSM candidates in a group

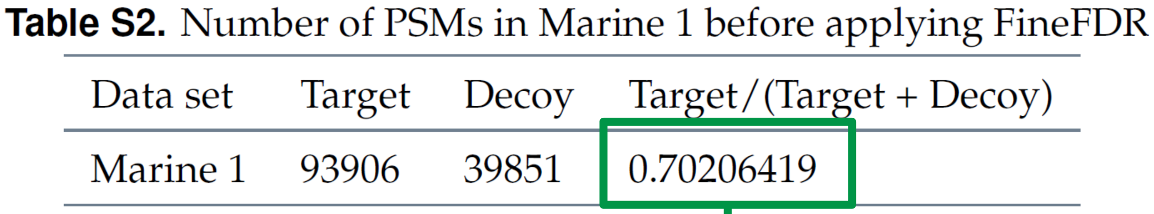


Table S2. Number of PSMs in Marine 1 before applying FineFDR

Data set	Target	Decoy	Target/(Target + Decoy)
Marine 1	93906	39851	0.70206419

The percentage of target PSM candidates in a random group without efficient grouping would be close to that in the original method without grouping.

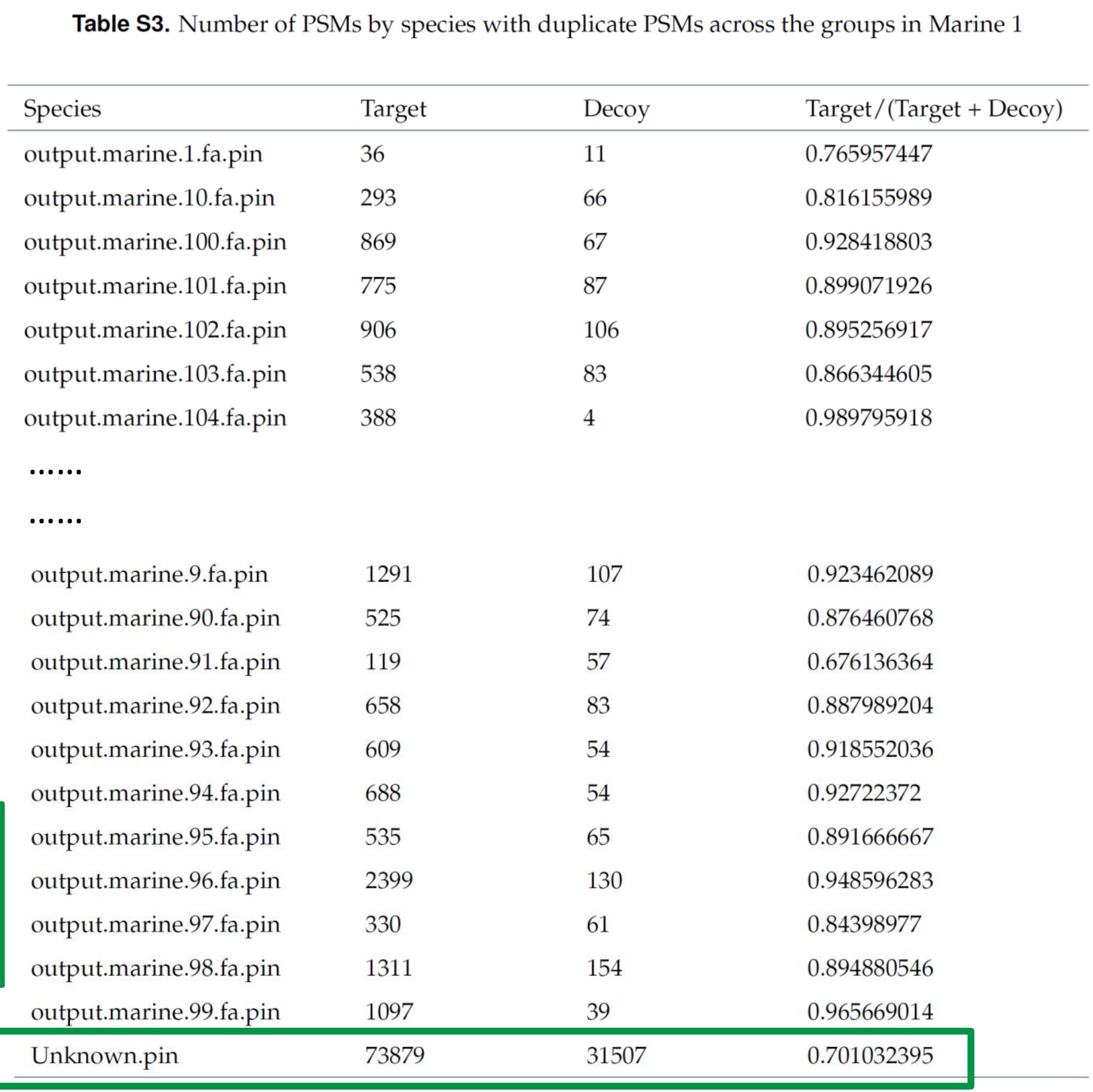


Table S3. Number of PSMs by species with duplicate PSMs across the groups in Marine 1

Species	Target	Decoy	Target/(Target + Decoy)
output.marine.1.fa.pin	36	11	0.765957447
output.marine.10.fa.pin	293	66	0.816155989
output.marine.100.fa.pin	869	67	0.928418803
output.marine.101.fa.pin	775	87	0.899071926
output.marine.102.fa.pin	906	106	0.895256917
output.marine.103.fa.pin	538	83	0.866344605
output.marine.104.fa.pin	388	4	0.989795918
......			
......			
output.marine.9.fa.pin	1291	107	0.923462089
output.marine.90.fa.pin	525	74	0.876460768
output.marine.91.fa.pin	119	57	0.676136364
output.marine.92.fa.pin	658	83	0.887989204
output.marine.93.fa.pin	609	54	0.918552036
output.marine.94.fa.pin	688	54	0.92722372
output.marine.95.fa.pin	535	65	0.891666667
output.marine.96.fa.pin	2399	130	0.948596283
output.marine.97.fa.pin	330	61	0.84398977
output.marine.98.fa.pin	1311	154	0.894880546
output.marine.99.fa.pin	1097	39	0.965669014
Unknown.pin	73879	31507	0.701032395

Conclusion

Contribution

- A novel FDR estimation framework, called FineFDR, was proposed for metaproteomics.
- FineFDR controls the FDR separately for PSMs/peptides/proteins from the different taxonomic units.
- FineFDR achieved higher precision and more PSM, peptide, and protein identifications.
- FineFDR is freely available under the GNU GPL license at https://github.com/Biocomputing-Research-Group/FDR.

Future Work

- FineFDR will support more search engines and post-search tools in future releases.
- Beyond Taxonomy-specific FDR control, we are investigating more techniques to mitigate the FDR estimation bias in metaproteomics.

Thank you for your time!

CORRESPONDENCE

xuan.guo@unt.edu

ACKNOWLEDGMENT

Research reported in this publication was supported by the National Library of Medicine of the National Institutes of Health, the National Center for Complementary & Integrative Health, and the National Institute of General Medical Sciences under the award numbers R15LM013460 and R01AT011618.

The authors acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing HPC resources that have contributed to the research results reported in this paper. URL: https://tacc.utexas.edu