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

1. ABSTRACT

Microbial community proteomics, also termed metaproteomics, investigates all proteins expressed by a microbiota. Tandem mass spectrometry (MS/MS) is the typical method for identifying proteins in metaproteomics, which involves searching the mass spectra against a protein sequence database. A major post-analysis step is controlling the false discovery rate (FDR), i.e., the ratio of false positives to the total number of annotations. The current popular target-decoy FDR estimation method treats all the peptides and proteins equally and overlooks that they could have varied probabilities of being identified. In this study, we report FineFDR, a framework for FDR assessment at fine-grained levels with taxonomy information considered. FineFDR groups the identified peptide-spectrum matches, peptides, and proteins from different taxonomic units and estimates the FDR in each group separately. Empirical experiments on the simulated and real-world data sets demonstrate that our FineFDR achieved higher precision and more peptide and protein identifications compared to the state-of-the-art methods, such as Comet, Percolator, TIDD, and Tailor. FineFDR is freely available under the GNU GPL license at https://github.com/Biocomputing-Research-Group/FDR.

2. SUMMARY OF RESULTS

Based on our previous study [1], Percolator [2], which is one of the most widely-used and adopted filters, outperformed other popular filtering algorithms, including Q-ranker [3], PeptideProphet [4], and iProphet [5]. Hence we picked Percolator[2] for comparison. On the Mock U1 [6] data set, FineFDR improved the identification rates of PSMs, peptides, and proteins by 4.0%, 3.0%, and 0.5% compared to the baseline method using Comet E-value; 0.1%, 0.3%, and 0.2% compared to the baseline method using Percolator p-score; 2.5%, 1.8%, and 0.3% compared to the baseline method using TIDD SVM_Prob; 6.0%, 4.4%, and 1.9% compared to the baseline method using Tailor score. For the marine communities, FineFDR averagely identified 12.5%, 13.4%, and 6.9% more PSMs, peptides, and proteins than the baseline method using Comet E-value; 1.7%, 1.6%, and 2.0% more PSMs, peptides, and proteins than the baseline method using Percolator p-score; 12.8%, 13.3%, and 8.7% more PSMs, peptides, and proteins than the baseline method using TIDD SVM_Prob; 23.1%, 23.2%, and 14.1% more PSMs, peptides, and proteins than the baseline method using Tailor score. For the soil communities, FineFDR averagely obtained 5.7%, 5.0%, and 3.9% more PSM, peptide, and protein than the baseline method using Comet E-value; 0.2%, 2.8%, and 3.4% more PSMs, peptides, and proteins than the baseline method using Percolator p-score; 2.5%, 2.8%, and 1.5% more PSMs, peptides, and proteins than the baseline method using TIDD SVM_Prob; 6.9%, 6.7%, and 6.5% more PSMs, peptides, and proteins than the baseline method using Tailor score. On the human gut data set, FineFDR boosted the identification rates of PSMs, peptides, and proteins by 6.3%, 8.1%, and 4.7% for the baseline method using Comet E-value; 9.6%, 3.0%, and 5.3% for the baseline method using Percolator p-score; 1.8%, 4.5%, and 2.9% for the baseline method using TIDD SVM_Prob; 1.8%, 4.9%, and 3.0% for the baseline method using Tailor score.

3. FIGURES AND TABLES

In U1, there are 29 known species which we used to build the taxonomy database. Fig. S1 illustrates the identification improvements at the species level. The species labels were obtained from the project [6] generating the Mock U1 data set.

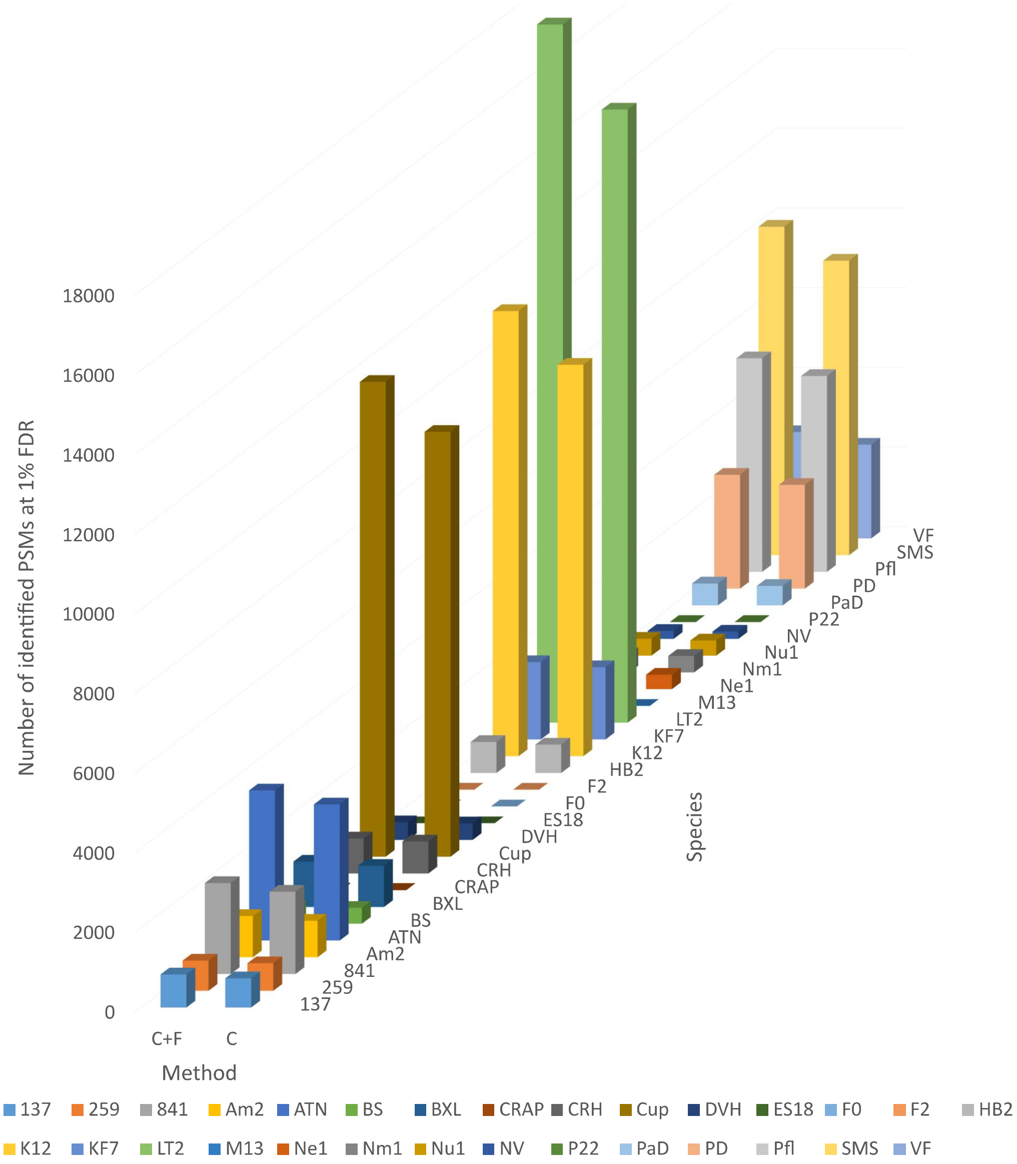

Fig. S1. PSM identification improvements by species for the Mock U1

Table S1 shows the average computational time on our platform using Comet with FineFDR.

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

Table S2 shows the statistical information of PSMs in the data set Marine 1. The method applied was Comet.

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

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

Table S3 shows the statistical information of PSMs by species with duplicate PSMs across the groups in the data set Marine 1. The method applied was Comet + FineFDR.

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.105.fa.pin	129	30	0.811320755
output.marine.106.fa.pin	1474	63	0.959011061
output.marine.107.fa.pin	60	9	0.869565217
output.marine.108.fa.pin	773	93	0.8926097
output.marine.109.fa.pin	1954	146	0.93047619
output.marine.11.fa.pin	325	94	0.775656325
output.marine.110.fa.pin	649	68	0.905160391
output.marine.111.fa.pin	20	5	0.8
output.marine.112.fa.pin	669	101	0.868831169
output.marine.113.fa.pin	362	22	0.942708333
output.marine.114.fa.pin	236	42	0.848920863
output.marine.115.fa.pin	127	44	0.742690058
output.marine.116.fa.pin	230	28	0.891472868
output.marine.117.fa.pin	1241	1053	0.54097646
output.marine.118.fa.pin	1226	109	0.91835206
output.marine.119.fa.pin	161	39	0.805
output.marine.12.fa.pin	439	50	0.897750511
output.marine.120.fa.pin	814	25	0.970202622
output.marine.121.fa.pin	100	9	0.917431193
output.marine.122.fa.pin	3	3	0.5

output.marine.123.fa.pin	180	27	0.869565217
output.marine.124.fa.pin	154	34	0.819148936
output.marine.125.fa.pin	111	14	0.888
output.marine.126.fa.pin	115	5	0.958333333
output.marine.127.fa.pin	597	80	0.88183161
output.marine.128.fa.pin	648	27	0.96
output.marine.129.fa.pin	793	68	0.921022067
output.marine.13.fa.pin	358	16	0.957219251
output.marine.130.fa.pin	276	64	0.811764706
output.marine.131.fa.pin	170	9	0.94972067
output.marine.132.fa.pin	2733	125	0.956263121
output.marine.133.fa.pin	216	10	0.955752212
output.marine.134.fa.pin	298	70	0.809782609
output.marine.135.fa.pin	1305	70	0.949090909
output.marine.136.fa.pin	42	8	0.84
output.marine.137.fa.pin	1363	73	0.949164345
output.marine.138.fa.pin	450	66	0.872093023
output.marine.139.fa.pin	400	93	0.811359026
output.marine.14.fa.pin	744	107	0.87426557
output.marine.140.fa.pin	1191	100	0.922540666
output.marine.141.fa.pin	79	16	0.831578947
output.marine.142.fa.pin	522	55	0.904679376
output.marine.143.fa.pin	340	14	0.960451977
output.marine.144.fa.pin	275	20	0.93220339
output.marine.145.fa.pin	38	7	0.844444444
output.marine.146.fa.pin	86	36	0.704918033
output.marine.147.fa.pin	69	49	0.584745763
output.marine.148.fa.pin	85	14	0.858585859
output.marine.149.fa.pin	289	55	0.840116279
output.marine.15.fa.pin	465	49	0.904669261
output.marine.150.fa.pin	101	38	0.726618705
output.marine.151.fa.pin	1261	39	0.97
output.marine.152.fa.pin	2445	501	0.8299389
output.marine.153.fa.pin	36	15	0.705882353
output.marine.154.fa.pin	151	16	0.904191617
output.marine.155.fa.pin	280	74	0.790960452
output.marine.156.fa.pin	2166	88	0.960958296
output.marine.157.fa.pin	208	100	0.675324675
output.marine.158.fa.pin	430	64	0.870445344

output.marine.159.fa.pin	58	13	0.816901408
output.marine.16.fa.pin	664	66	0.909589041
output.marine.160.fa.pin	1093	84	0.928632116
output.marine.161.fa.pin	82	13	0.863157895
output.marine.162.fa.pin	985	56	0.946205572
output.marine.163.fa.pin	226	31	0.879377432
output.marine.164.fa.pin	181	64	0.73877551
output.marine.165.fa.pin	173	76	0.694779116
output.marine.166.fa.pin	199	199	0.5
output.marine.167.fa.pin	113	20	0.84962406
output.marine.168.fa.pin	36	7	0.837209302
output.marine.169.fa.pin	794	127	0.862106406
output.marine.17.fa.pin	380	56	0.871559633
output.marine.18.fa.pin	529	54	0.907375643
output.marine.19.fa.pin	19	17	0.527777778
output.marine.2.fa.pin	552	105	0.840182648
output.marine.20.fa.pin	821	74	0.917318436
output.marine.21.fa.pin	45	11	0.803571429
output.marine.22.fa.pin	525	52	0.909878683
output.marine.23.fa.pin	463	11	0.976793249
output.marine.24.fa.pin	789	21	0.974074074
output.marine.25.fa.pin	356	51	0.874692875
output.marine.26.fa.pin	349	80	0.813519814
output.marine.27.fa.pin	43	7	0.86
output.marine.28.fa.pin	165	11	0.9375
output.marine.29.fa.pin	780	74	0.913348946
output.marine.3.fa.pin	788	63	0.925969448
output.marine.30.fa.pin	152	14	0.915662651
output.marine.31.fa.pin	69	4	0.945205479
output.marine.32.fa.pin	591	82	0.878157504
output.marine.33.fa.pin	178	75	0.703557312
output.marine.34.fa.pin	33	10	0.76744186
output.marine.35.fa.pin	50	14	0.78125
output.marine.36.fa.pin	679	109	0.861675127
output.marine.37.fa.pin	1579	110	0.934872706
output.marine.38.fa.pin	285	122	0.7002457
output.marine.39.fa.pin	1033	212	0.829718876
output.marine.4.fa.pin	1469	59	0.961387435
output.marine.40.fa.pin	30	11	0.731707317

output.marine.41.fa.pin	1342	72	0.949080622
output.marine.42.fa.pin	407	33	0.925
output.marine.43.fa.pin	775	59	0.929256595
output.marine.44.fa.pin	42	14	0.75
output.marine.45.fa.pin	38	6	0.863636364
output.marine.46.fa.pin	950	190	0.833333333
output.marine.47.fa.pin	153	31	0.831521739
output.marine.48.fa.pin	41	14	0.745454545
output.marine.49.fa.pin	588	91	0.865979381
output.marine.5.fa.pin	808	39	0.953955136
output.marine.50.fa.pin	503	63	0.88869258
output.marine.51.fa.pin	44	13	0.771929825
output.marine.52.fa.pin	458	97	0.825225225
output.marine.53.fa.pin	686	59	0.920805369
output.marine.54.fa.pin	806	77	0.912797282
output.marine.55.fa.pin	491	86	0.850953206
output.marine.56.fa.pin	131	13	0.909722222
output.marine.57.fa.pin	458	45	0.910536779
output.marine.58.fa.pin	378	20	0.949748744
output.marine.59.fa.pin	243	23	0.913533835
output.marine.6.fa.pin	200	69	0.743494424
output.marine.60.fa.pin	927	56	0.943031536
output.marine.61.fa.pin	453	60	0.883040936
output.marine.62.fa.pin	685	36	0.950069348
output.marine.63.fa.pin	35	18	0.660377358
output.marine.64.fa.pin	48	8	0.857142857
output.marine.65.fa.pin	104	15	0.87394958
output.marine.66.fa.pin	336	34	0.908108108
output.marine.67.fa.pin	188	17	0.917073171
output.marine.68.fa.pin	177	67	0.725409836
output.marine.69.fa.pin	3675	351	0.912816692
output.marine.7.fa.pin	807	101	0.88876652
output.marine.70.fa.pin	74	13	0.850574713
output.marine.71.fa.pin	161	33	0.829896907
output.marine.72.fa.pin	2227	149	0.937289562
output.marine.73.fa.pin	297	39	0.883928571
output.marine.74.fa.pin	677	54	0.926128591
output.marine.75.fa.pin	269	14	0.950530035
output.marine.76.fa.pin	1699	114	0.937120794

output.marine.77.fa.pin	79	23	0.774509804
output.marine.78.fa.pin	410	89	0.821643287
output.marine.79.fa.pin	35	4	0.897435897
output.marine.8.fa.pin	44	9	0.830188679
output.marine.80.fa.pin	1127	64	0.946263644
output.marine.81.fa.pin	164	7	0.959064327
output.marine.82.fa.pin	352	50	0.875621891
output.marine.83.fa.pin	757	94	0.889541716
output.marine.84.fa.pin	564	99	0.850678733
output.marine.85.fa.pin	225	33	0.872093023
output.marine.86.fa.pin	588	58	0.910216718
output.marine.87.fa.pin	2238	162	0.9325
output.marine.88.fa.pin	12	7	0.631578947
output.marine.89.fa.pin	696	65	0.914586071
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

Compared to the original method (Comet), the taxonomy-specific method (Comet + FineFDR), which distinguish PSM candidates based on taxonomic groups, shows the power to promote the percentage of target PSM candidate in a group. And we can assume the percentage of target PSM candidates in a random group without efficient grouping would be close to the original method without grouping.

REFERENCES

1. S. Feng, R. Sterzenbach, and X. Guo, "Deep learning for peptide identification from metaproteomics datasets," J. Proteomics **247**, 104316 (2021).
2. M. J. MacCoss, W. S. Noble, L. Käll *et al.*, "Fast and accurate protein false discovery rates on large-scale proteomics data sets with percolator 3.0," J. Am. Soc. for Mass Spectrom. **27**, 1719–1727 (2016).
3. M. Spivak, J. Weston, L. Bottou, L. Kall, and W. S. Noble, "Improvements to the percolator algorithm for peptide identification from shotgun proteomics data sets," J. proteome research **8**, 3737–3745 (2009).
4. A. Keller, A. I. Nesvizhskii, E. Kolker, and R. Aebersold, "Empirical statistical model to estimate the accuracy of peptide identifications made by ms/ms and database search," Anal. chemistry **74**, 5383–5392 (2002).
5. D. Shteynberg, E. W. Deutsch, H. Lam, J. K. Eng, Z. Sun, N. Tasman, L. Mendoza, R. L. Moritz, R. Aebersold, and A. I. Nesvizhskii, "iprophet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates," Mol. & cellular proteomics **10** (2011).
6. D. Estève, N. Boulet, C. Belles, A. Zakaroff-Girard, P. Decaunes, A. Briot, Y. Veeranagouda, M. Didier, A. Remaury, J. Guillemot *et al.*, "Lobular architecture of human adipose tissue defines the niche and fate of progenitor cells," Nat. communications **10**, 1–16 (2019).